Computational Privacy: Towards Privacy-Conscientious Uses of Metadata

by

Yves-Alexandre de Montjoye

M.Sc. Ingénieur civil en mathématiques appliquées, Université catholique de Louvain

M.Sc. Wiskundige Ingenieurstechnieken, Katholieke Universiteit Leuven

Ingénieur des Arts et Manufactures (M.Sc.), Ecole Centrale Paris

B.Sc. Ingénieur Civil, Université catholique de Louvain

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2015

Signature redacted

Author .....
Program in Media Arts and Sciences, School of Architecture and Planning
August 07, 2015

Certified by ..... Signature redacted .....
Prof. Alex "Sandy" Pentland
Toshiba Professor of Media Arts and Sciences
Thesis Supervisor

Signature redacted

Accepted by .....
Prof. Pattie Maes
Academic Head
Program in Media Arts and Sciences

Computational Privacy: Towards Privacy-Conscientious Uses of Metadata

by

Yves-Alexandre de Montjoye

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
on August 07, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The breadcrumbs left behind by our technologies have the power to fundamentally transform the health and development of societies. Metadata about our whereabouts, social lives, preferences, and finances can be used for good but can also be abused. In this thesis, I show that the richness of today's datasets have rendered traditional data protections strategies outdated, requiring us to deeply rethink our approach.

First, I show that the concept of anonymization, central to legal and technical data protection frameworks, does not scale. I introduce the concept of *unicity* to study the risks of re-identification of large-scale metadata datasets given p points. I then use *unicity* to show that four spatio-temporal points are enough to uniquely identify 95% of people in a mobile phone dataset and 90% of people in a credit card dataset. In both cases, I also show that traditional de-identification strategies such as data generalization are not sufficient to approach anonymity in modern high-dimensional datasets.

Second, I argue that the second pillar of data protection, risk assessment, is similarly crumbling as data gets richer. I show, for instance, how standard mobile phone data–information on how and when somebody calls or texts–can be used to predict personality traits up to 1.7 times better than random. The risk of inference in big data will render comprehensive risks assessments increasingly difficult and, moving forward, potentially irrelevant as they will require evaluating what can be inferred now, and in the future, from rich data.

However, this data has a great potential for good especially in developing countries. While it is highly unlikely that we will ever find a magic bullet or even a one-size-fits-all approach to data protection, there are ways that exist to use metadata in privacy-conscientious ways. I finish this thesis by discussing technical solutions (incl. privacy-through-security ones) which, when combined with legal and regulatory frameworks, provide a reasonable balance between the imperative of using this data and the legitimate concerns of the individual and society.

Thesis Supervisor: Prof. Alex "Sandy" Pentland
Title: Toshiba Professor of Media Arts and Sciences

Computational Privacy: Towards Privacy-Conscientious Uses of Metadata

by

Yves-Alexandre de Montjoye

This doctoral thesis has been examined by a Committee as follows:

Professor Alex "Sandy" Pentland ..... Signature redacted

Thesis Supervisor

Toshiba Professor of Media Arts and Sciences

Massachusetts Institute of Technology

Computational Privacy: Towards Privacy-Conscientious Uses of Metadata

by

Yves-Alexandre de Montjoye

This doctoral thesis has been examined by a Committee as follows:

Signature redacted

Professor Gary King .....

Member, Thesis Committee

Albert J. Weatherhead III University Professor

Harvard University

Computational Privacy: Towards Privacy-Conscientious Uses of Metadata

by

Yves-Alexandre de Montjoye

This doctoral thesis has been examined by a Committee as follows:

Signature redacted

Professor Alessandro Acquisti .....

Member, Thesis Committee

Professor of Information Technology and Public Policy

Carnegie Mellon University

Acknowledgments

First and foremost, my advisor Alex "Sandy" Pentland for pushing me to think further, to focus on impact, all while always giving me his unconditional support. My committee Gary King and Alessandro Acquisti for their feedback, support, and encouragements. It has been a fantastic journey and I am looking forward to continue our collaboration. César Hidalgo, Cameron Kerry, Jake Kendall, Linus Bengtsson, Latanya Sweeney, Julien Hendrickx for everything they taught me about science, policy, and humanitarian affairs. Vincent Blondel, Nathan Eagle, Aaron Clauset, and Luis Bettencourt for giving me the love and deep appreciation for the beauty of research. Joost Bonsen and John Clippinger for always encouraging me to think out of the box. My fellow labbers and collaborators: Coco Krumme, Arek Stopczynski, Brian Sweatt, Laura Radaelli, Luc Rocher, Florent Robic, Erez Shmueli, Sune Lehmann, Iyad Rahwan, Gordon Wetzstein, Christoph Stadtfeld, Scott Greenwald, Eaman Jahani, Vivek Singh, Ben Waber, Pål Sundsøy, Johannes Bjelland, Riley Crane, Jeff Schmitz, Emmanuel Letouzé, Manuel Cebrian, Bruno Lepri, Oren Lederman, Dhaval Adjodah, Abdullah Almaatouq, Peter Krafft, Alejandro Noriega Campero, and, of course, the irreplaceable Nicole Freedman. Jean, Jean-Benoît, and the Cambridge crew: Thomas, Daan, Joris, Michiel, Maxime, Phebe, Jordi, Martin, Pierre, Emilia, and Lisa. Final acknowledgments go to my family: my parents Yves and Carol, my sisters Laurence and Stéphanie, my brother-in-law Julien, and the little Raphaël.

Contents

Introduction

Metadata have the power to fundamentally transform the health and development of societies. The breadcrumbs left behind by the technologies of our daily lives dramatically increase our capacity to measure and understand the behavior of individuals and societies. The recent availability of large-scale behavioral datasets has been compared by researchers to the invention of the microscope [206] and has given rise to a new field, computational social science [129]. Metadata have great potential beyond research. For commercial purposes, but also as a major new source of information in developing countries [132]. For instance, metadata datasets have already been used to generate accurate population censuses [89], to infer linguistic and ethnic boundaries [57], to compare lifestyles in rural and urban areas [93], and to help improve responses to natural disasters [51].

The amount of metadata collected and their scope will only increase. More than six billion mobile phones in the world are already generating metadata, including locational information, every time a phone call is made or a text is sent. Vehicle tracking GPSs cost less than $40 and E-ZPass records more than 2.6 billion vehicles crossings bridges, tunnels, and highways every year [92]. In the United States, one hundred and twenty billion non-cash payments are been processed and recorded annually [28]. In Kenya, 31% of the GDP is accounted for by transactions made on Safaricom's M-Pesa [142]. Worldwide, 39% of the population is using the internet [112], Americans for more than 25 hours a week [158]. Moving forward, an ever increasing fraction of our daily activities will be recorded in metadata by new sensors such as wearables or the Internet of Things.

These metadata capture the most intimate details of our lives: rich information

about our whereabouts, social life, preferences, and finances. Such sensitive and personal information can be be used for good but can also be abused [33, 145]. Privacy has been foundational to the development of our societies [122]. As new data collection and analysis techniques are developed and deployed, it is essential to ensure that our legal and technical approaches to data protection keep up with technology. There are obvious benefits to the use of metadata datasets, but this first requires a solid quantitative understanding of their privacy. Such understanding will enable us to truly find the right balance between the privacy challenges of metadata and their great potential for good. Only then, will we be able to use this data in privacy-conscientious ways and in accordance with our choice of society.

In the first chapter of this thesis, I will show what I call the limits of anonymization of modern high-dimensional datasets. I will introduce *unicity* and, using mobile phones and credit cards metadata, I will show how individuals can easily be re-identified in simply anonymized datasets and how traditional de-identification strategies, such as data generalization, are not enough to provide anonymity.

In the second chapter of this thesis, I will discuss what I call the risk of inference of rich datasets and argue that it will render comprehensive risks assessments increasingly difficult. Here, I will show, for instance, how standard mobile phone metadata can be combined with machine learning algorithms to predict personality traits up to 1.7 times better than random.

Finally, in the third chapter of this thesis, I will argue that the limits of anonymization and the risk of inference require us to rethink our approach to data protection. I will then discuss technical solutions for the privacy-conscientious use of metadata with a focus on privacy-through-security frameworks.

Chapter 1

The Limits of Anonymization

The notion of anonymity has long been central to finding the balance between the utility of the data and its privacy: the so-called privacy-utility trade-off. While anonymity, from the Greek words *an* and *onoma*, can be translated literally to "without name", its current understanding is that the data cannot be re-identified. The risk of privacy loss or of the data being abused is strongly reduced if the data cannot be linked back to an individual.

While anonymity had historically been achieved through the removal of names and obvious identifiers, it became clear at the beginning of the century that the mere absence of direct identifiers (pseudonymous data) might not be enough to prevent individuals from being singled out in the data and re-identified. For instance, Latanya Sweeney at Carnegie Mellon University showed that 87% of the U.S. population is uniquely identified by their date of birth, gender, and 5-digit zip code [192]. This work and others gave rise to the notion of quasi-identifiers. Quasi-identifiers are pieces of information which, although they do not directly identify an individual, could be collected and combined by an attacker to re-identify an individual in a dataset.

Since then, the concept of quasi-identifiers has been central to legal and technical privacy work. From a technical perspective, k-anonymity has been developed to ensure that no combination of quasi-identifiers could be associated with less than k individuals. k-anonymity is achieved through generalization of records–for example, by releasing the year of birth instead of the full date–and through suppression of either columns or records. k-anonymity has been shown to be NP-hard [149], but good approximations can be found [49]. Some of the potential limitations of k-anonymity have, furthermore, been addressed by subsequent metrics. For instance, l-diversity aims at maintaining the diversity of sensitive fields [141], while t-closeness takes into account the distribution of sensitive attributes in each class [134]. From a legal perspective, the concept of Personally Identifiable Information (PII) is similar to quasi-identifiers, and is at the basis of most privacy regulation.

Results from Sections 1.1 and 1.2, however, put into question the achievability of meaningful anonymity and provable de-identification of large-scale metadata datasets. Taken together, they challenge our reliance on anonymization-based solutions as one

of the primary approaches to data protection.

1.1 Mobile Phone Data[1]

Derived 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 [122]. 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 [69]. 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 [205].

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. 1-1A-B] have been used in the past for research purposes [117, 44, 70, 95, 94, 93, 118, 156, 160, 196, 172, 43, 148, 108, 185] and to provide personalized services to users [12]. 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 [59]. These include the movements of a competitor sales force, attendance of a particular church or an individual's presence

[1]Published as de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. and Blondel, V.D. Unique in the Crowd: The privacy bounds of human mobility. Nature S.Rep. 3, 1376; DOI:10.1038/srep01376 (2013).

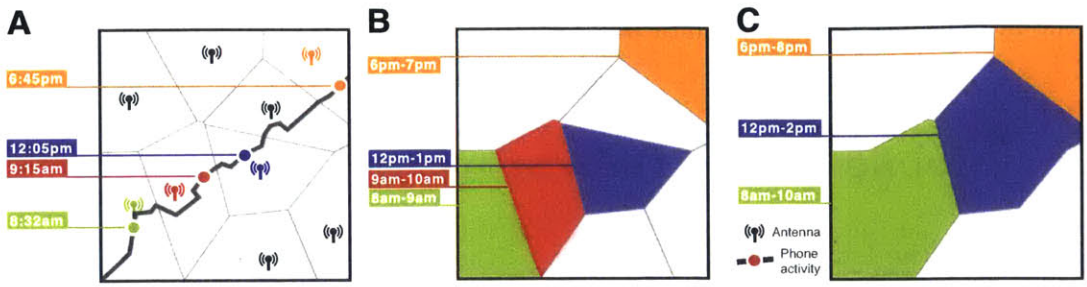


Figure 1-1: **(A)** Trace of an anonymized mobile phone user during a day. The dots represent the times and locations where the user made or received a call. Every time the user has such an interaction, the closest antenna that routes the call is recorded. **(B)** The same user's trace as recorded in 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 8:32am and 9:15am interactions indistinguishable.

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" [2]. 65.5B geo-tagged payments are made per year in the US [174] while Skyhook wireless is resolving 400M user's WiFi location every day [26]. Furthermore, it is estimated that a third of the 25B copies of applications available on Apple's App Store access a user's geographic location [1, 3], and that the geo-location of ~50% of all iOS and Android traffic is available to ad networks [19]. 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 the health record of the governor of Massachusetts [194]. In another, mobile phone data

have been re-identified using users' top locations [215]. Finally, part of the Netflix challenge dataset was re-identified using outside information from The Internet Movie Database [155].

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.

1.1.1 Uniqueness of Human Mobility

In 1930, Edmond Locard showed that 12 points are needed to uniquely identify a fingerprint [136]. 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 $\mathcal{E}$, the uniqueness of traces, by extracting from D the subset of trajectories $\mathrm{S}(I_p)$ that match the p points composing I_p [See Materials]. A trace is unique if $|\mathrm{S}(I_p)| = 1$, containing only one trace. For example, in Fig. 1-2A, 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

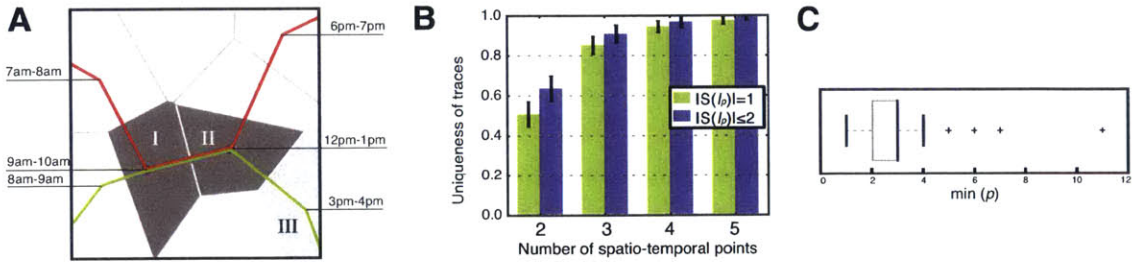


Figure 1-2: **(A)** $I_{p=2}$ means that the information available to the attacker consist of two 7am-8am spatio-temporal points (I and II). In this case, the target was in zone I between 9am to 10am and in zone II between 12pm to 1pm. In this example, the traces of two anonymized users (red and green) are compatible with the constraints defined by $I_{p=2}$. The subset $\mathrm{S}(I_{p=2})$ 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. $|\mathrm{S}(I_p)| = 1$. The blue bars represent the fraction of $|\mathrm{S}(I_p)| \leq 2$. 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 1.5M 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.

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 1.5M 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® [6]. 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. 1-3]. 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, $\sim$ 2000 inhabitants each, covering areas ranging from 0.15 km^2 in cities to 15 km^2 in rural areas. The number of antennas is strongly correlated with population density ($R^2 = .6426$) [Fig. 1-3C]. The same is expected from businesses, places in location-based social networks, or WiFi hotspots.

Fig. 1-2B shows the fraction of unique traces ($\mathcal{E}$) as a function of the number of available points p. Four randomly chosen points are enough to uniquely characterize 95% of the users ($\mathcal{E} > .95$), whereas two randomly chosen points still uniquely characterize more than 50% of the users ($\mathcal{E} > .5$). This shows that mobility traces are highly unique, and can therefore be re-identified using little outside information.

1.1.2 Scaling Properties

Nonetheless, $\mathcal{E}$ 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 1-1C]. 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

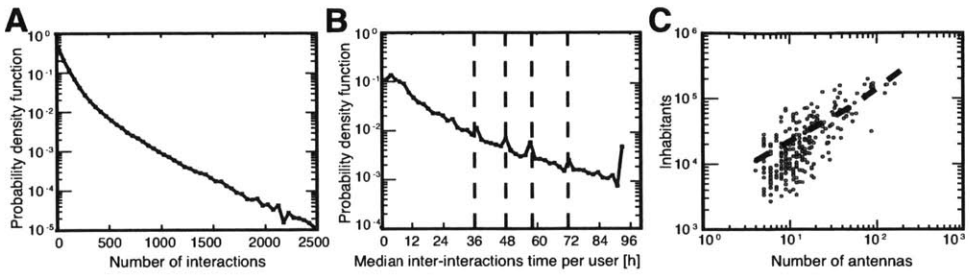


Figure 1-3: **(A)** Probability density function of the amount of recorded spatio-temporal points per user during a month. **(B)** Probability density function of the median inter-interaction time with the service. **(C)** The number of antennas per region is correlated with its population (R^2=.6426). These plots strongly emphasize the discrete character of our dataset and its similarities with datasets such as the one collected by smartphone apps.

Materials]. Both of these aggregations are bound to decrease $\mathcal{E}$, and therefore, make re-identification harder.

Fig. 1-4A shows how the uniqueness of mobility traces $\mathcal{E}$ depends on the spatial and temporal resolution of the data. This reduction, however, is quite gradual. Given four points (p=4), we find that $\mathcal{E} > .5$ 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, $\mathcal{E} > .6$ in a dataset with a temporal resolution of $h = 15$ hours or a spatial resolution of $v = 15$ antennas while $\mathcal{E} < .4$ in a dataset with a temporal resolution of $h = 7$ hours and a spatial resolution of $v = 7$ antennas [Fig. 1-4A].

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. 1-4B and C show that the uniqueness of a trace decreases as the power function $\mathcal{E} = \alpha - x^{\beta}$, for decreases in both the spatial and temporal resolution (x), and for all considered $p = 4$, 6, 8 and 10 (see Table 1.1.5). The uniqueness of human mobility can thus be expressed using the single formula: $\mathcal{E} = \alpha - (v * h)^{\beta}$. We find that this power function fits the data better than other

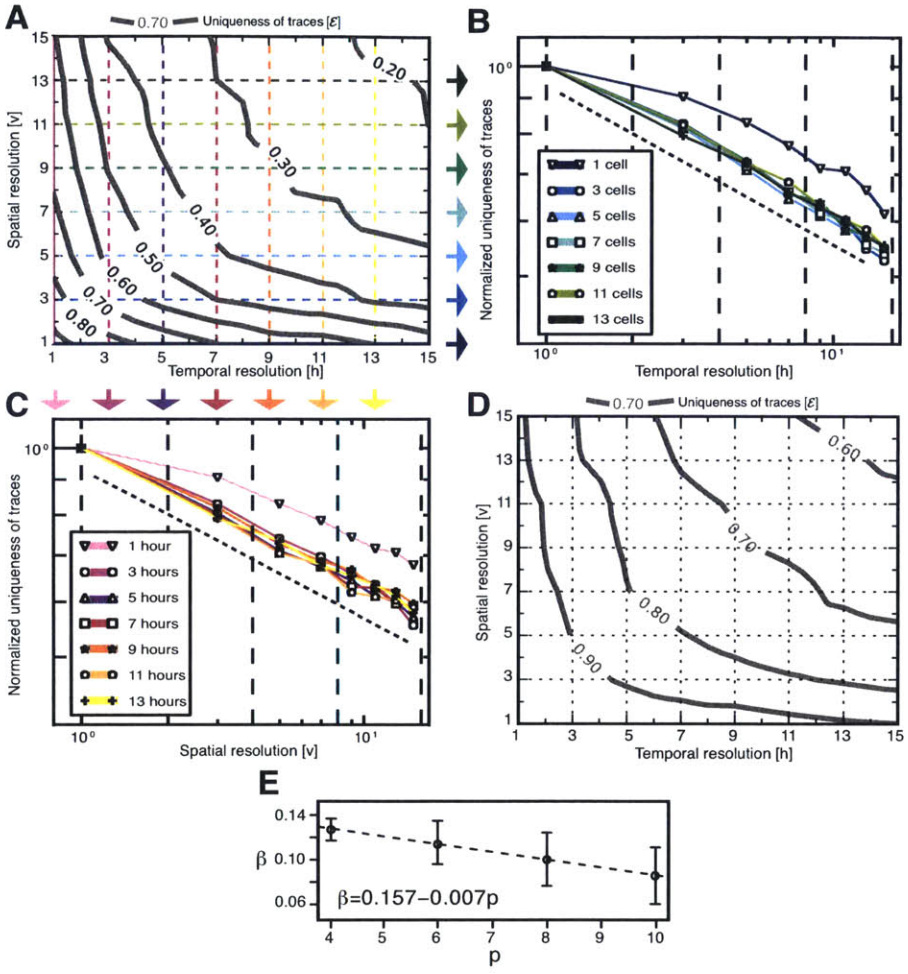


Figure 1-4: Uniqueness of traces [$\mathcal{E}$] when we lower the resolution of the dataset with **(A)** $p = 4$ and **(D)** $p = 10$ points. It is easier to attack a dataset that is coarse on one dimension and fine along another than a medium-grained dataset along both dimensions. Given four spatio-temporal 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 km^2 in urban areas to 15 km^2 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 $\mathcal{E} = \alpha - x^{\beta}$. **(E)** While $\mathcal{E}$ decreases according to a power function, its exponent β decreases linearly with the number of points p. Accordingly, it would always be possible to re-identify mobility datasets using information only on a few outside locations.

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 1.1.5). Both estimators for α and β are highly significant ($p < 0.001$) [48], 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. 1.1.5].

The power-law dependency of $\mathcal{E}$ 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. 1-2B shows that, as expected, $\mathcal{E}$ increases with p. The mitigating effect of p on $\mathcal{E}$ is mediated by the exponent β which decays linearly with p: $\beta = 0.157 - 0.007 * p$ [Fig. 1-4E]. 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 1.1.5).

Because of the functional of $\mathcal{E}$ on p through the exponent β, mobility datasets are likely to be re-identifiable using information on only a few outside locations.

1.1.3 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 $\mathcal{E}$. 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 [108]. 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 down-town Boston on a Friday evening.

This study is likely to underestimate $\mathcal{E}$, 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 entropy is 1.74, less than two locations [185]. 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. 1.1.5] 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 identified uniquely knowing only 11 locations [Fig. 1-2C].

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 workplace or travels abroad [155, 106]. 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 $\mathcal{E} \sim (v * h)^{\beta}$, $\beta \sim -p/100$. 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 [143] and knowing the bounds of individual's privacy will be crucial in the design of both future policies and information technologies.

1.1.4 Methods

The dataset

This work was performed using an anonymized mobile phone dataset that contains call information for ~1.5M 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. 1-1A]. The dataset's intrinsic spatial resolution is thus the maximal half-distance between antennas. The dataset's intrinsic temporal resolution is one hour [Fig. 1-1B].

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. 1-2A]. Note that this notion of compatibility can easily be extended to noisy or richer data. A brute force characterization is performed by extracting from the entire dataset of 1.5M users $\mathrm{S}(I_p)$, 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: $|\mathrm{S}(I_p)| \leq x$. A trace is uniquely characterized if the set contains exactly one trace: $|\mathrm{S}(I_p)| = 1$. 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. 1-2B shows that $.95 < \mathcal{E} < 1$ given $I_{p=4}$. Fig. 1-2C 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 [111]. Fig. 1.1.5 shows the resulting group size histogram optimized for clusters of size 4. Once antennas are aggregated into groups, their associated regions are merged.

1.1.5 Appendix

Fig. S1. Probability of having a location recorded per hour (blue, right axis) and per day (orange, left axis). Intuitively, knowing a point at 3AM is more likely to make a trace unique than a point at 4PM. As the given spatio-temporal points are drawn at random, they follow this temporal distribution. This makes our work a rather adversary case.

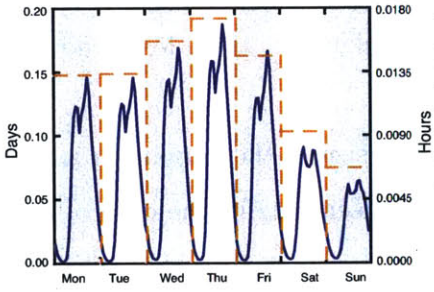
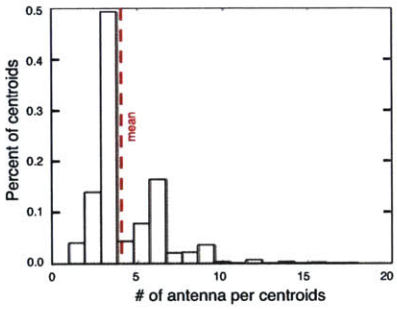


Fig. S2. Number of antenna per centroids when the algorithm for spatial aggregation aims at clusters of size four.



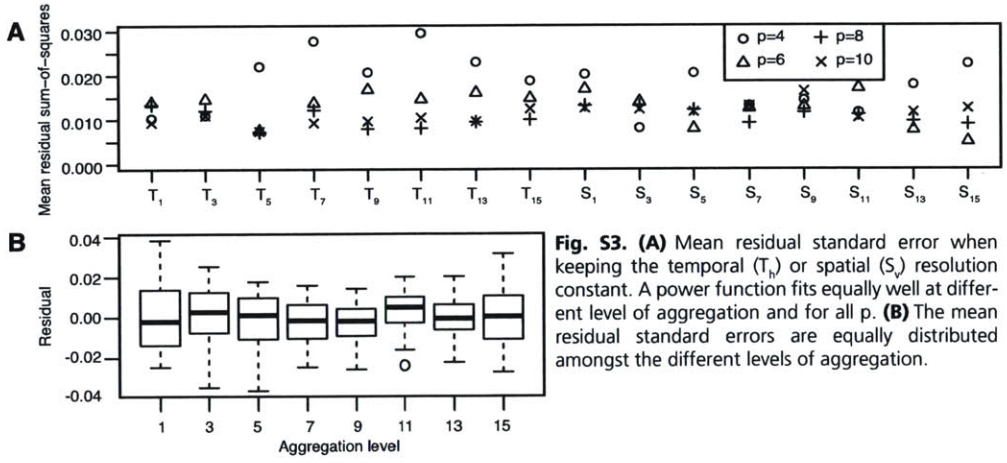


Fig. S3. (A) Mean residual standard error when keeping the temporal (T_h) or spatial (S_v) resolution constant. A power function fits equally well at different level of aggregation and for all p. **(B)** The mean residual standard errors are equally distributed amongst the different levels of aggregation.

Table S1. Power-law function fitting

	Parameters and goodness of fit of $\mathcal{E} = \alpha - x^{\beta}$			Goodness of fit of alternative functions		
				$\mathcal{E} = \alpha - \exp(\beta x)$	$\mathcal{E} = \alpha - \beta x$	$\mathcal{E} = \alpha - \exp x^{\beta}$
p	$< \beta >^{\dagger}$	MSRE	$<$ pseudo-R^2 $>$	$<$ pseudo-R^2 $>$	$<$ pseudo-R^2 $>$	$<$ pseudo-R^2 $>$
4	0.1282 +/- 0.009 ***	0.018	0.983	0.813 ‡	0.842 ‡	0.987
6	0.1164 +/- 0.019 ***	0.012	0.987	0.863 ‡	0.886 ‡	0.976 ‡
8	0.1011 +/- 0.024 ***	0.010	0.984	0.903 ‡	0.921 ‡	0.967 ‡
10	0.0860 +/- 0.025 ***	0.011	0.975	0.915 ‡	0.930 ‡	0.960 ‡
Overall one-tailed paired t-test between the MSRE:				p<0.001	p<0.001	p<0.001

†+/- as SD, *** indicates a p<0.001, ‡ Indicates a p<0.001 on a one-tailed paired t-test between the MSRE of $\mathcal{E} = \alpha - x^{\beta}$ and of alternative functions

1.2 Credit Card Data[2]

1.2.1 Introduction

Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities, or perform research. Ubiquitous technologies create personal metadata on a very large scale. Our smartphones, browsers, cars, or credit cards generate information about where we are, whom we call, or how much we spend. Scientists have compared this recent availability of large-scale behavioral data sets to the invention of the microscope [176]. New fields such as computational social science [129, 104, 207] rely on metadata to address crucial questions such as fighting malaria, studying the spread of information, or monitoring poverty [209, 67, 94]. The same metadata data sets are also used by organizations and governments. For example, Netflix uses viewing patterns to recommend movies, whereas Google uses location data to provide real-time traffic information, allowing drivers to reduce fuel consumption and time spent traveling [162].

The transformational potential of metadata data sets is, however, conditional on their wide availability. In science, it is essential for the data to be available and shareable. Sharing data allows scientists to build on previous work, replicate results, or propose alternative hypotheses and models. Several publishers and funding agencies now require experimental data to be publicly available [60, 147, 58]. Governments and businesses are similarly realizing the benefits of open data [15]. For example, Boston's transportation authority makes the real-time position of all public rail vehicles available through a public interface [40], whereas Orange Group and its subsidiaries make large samples of mobile phone data from Côte d'Ivoire and Senegal available to selected researchers through their Data for Development challenges [84, 56]. These metadata are generated by our use of technology and, hence, may reveal a lot about an individual [145, 82]. Making these data sets broadly avail-

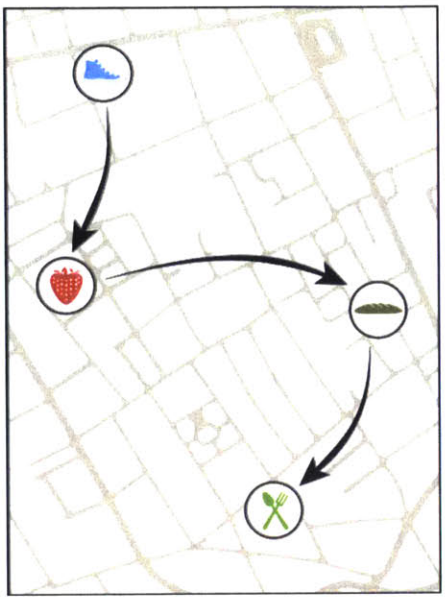
[2]Published as de Montjoye Y.-A., Radaelli L., Singh V. K., Pentland A. S., Unique in the shopping mall: On the reidentifiability of credit card metadata. Science 347 (6221), 536-539. DOI:10.1126/science.1256297 (2015).

able, therefore, requires solid quantitative guarantees on the risk of reidentification. A data set's lack of names, home addresses, phone numbers, or other obvious identifiers [such as required, for instance, under the U.S. personally identifiable information (PII) "specific-types" approach [178]], does not make it anonymous nor safe to release to the public and to third parties. The privacy of such simply anonymized data sets has been compromised before [79, 155, 184, 194].

Unicity quantifies the intrinsic reidentification risk of a data set [79]. It was recently used to show that individuals in a simply anonymized mobile phone data set are reidentifiable from only four pieces of outside information. Outside information could be a tweet that positions a user at an approximate time for a mobility data set or a publicly available movie review for the Netflix data set [155]. Unicity quantifies how much outside information one would need, on average, to reidentify a specific and known user in a simply anonymized data set. The higher a data set's unicity is, the more reidentifiable it is. It consequently also quantifies the ease with which a simply anonymized data set could be merged with another.

Financial data that include noncash and digital payments contain rich metadata on individuals' behavior. About 60% of payments in the United States are made using credit cards [28], and mobile payments are estimated to soon top $1 billion in the United States [97]. A recent survey shows that financial and credit card data sets are considered the most sensitive personal data worldwide [29]. Among Americans, 87% consider credit card data as moderately or extremely private, whereas only 68% consider health and genetic information private, and 62% consider location data private. At the same time, financial data sets have been used extensively for credit scoring [120], fraud detection [55], and understanding the predictability of shopping patterns [127]. Financial metadata have great potential, but they are also personal and highly sensitive. There are obvious benefits to having metadata data sets broadly available, but this first requires a solid understanding of their privacy.

To provide a quantitative assessment of the likelihood of identification from financial data, we used a data set D of 3 months of credit card transactions for 1.1 million users in 10,000 shops in an Organisation for Economic Co-operation and Development



shop	user_id	time	price	price_bin
	7abc1a23	09/23	$97.30	$49 – $146
	7abc1a23	09/23	$15.13	$5 – $16
	3092fc10	09/23	$43.78	$16 – $49
	7abc1a23	09/23	$4.33	$2 – $5
	4c7af72a	09/23	$12.29	$5 – $16
	89c0829c	09/24	$3.66	$2 – $5
	7abc1a23	09/24	$35.81	$16 – $49

Figure 1-5: Financial traces in a simply anonymized data set such as the one we use for this work. Arrows represent the temporal sequence of transactions for user 7abc1a23 and the prices are grouped in bins of increasing size.

country (Fig. 1-5). The data set was simply anonymized, which means that it did not contain any names, account numbers, or obvious identifiers. Each transaction was time-stamped with a resolution of 1 day and associated with one shop. Shops are distributed throughout the country, and the number of shops in a district scales with population density ($r^2 = 0.51, P < 0.001$) (Fig. 1-10).

1.2.2 Results

We quantified the risk of reidentification of D by means of unicity $\mathcal{E}$ [79]. Unicity is the risk of reidentification knowing p pieces of outside information about a user. We evaluate $\mathcal{E}_p$ of D as the percentage of its users who are reidentified with p randomly selected points from their financial trace. For each user, we extracted the subset $S(I_p)$ of traces that match the p known points (I_p). A user was considered reidentified in this correlation attack if $|S(I_p)| = 1$.

For example, let's say that we are searching for Scott in a simply anonymized credit card data set (Fig. 1-5). We know two points about Scott: he went to the

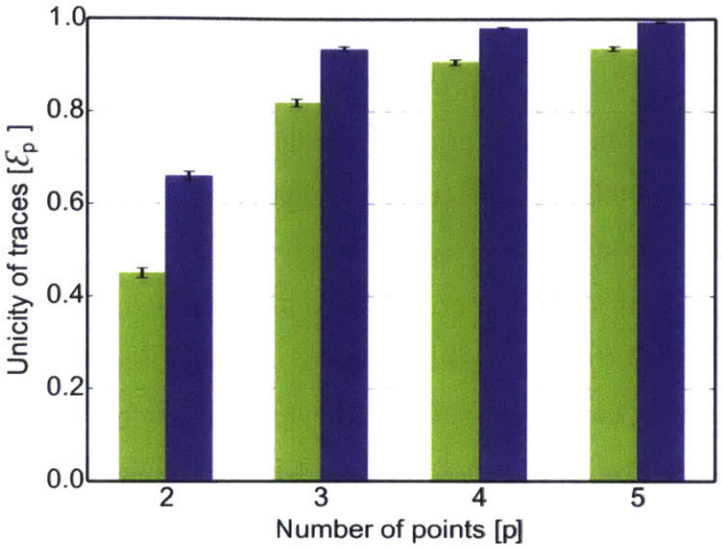


Figure 1-6: The unicity $\mathcal{E}$of the credit card data set given p points. The green bars represent unicity when spatiotemporal tuples are known. This shows that four spatiotemporal points taken at random ($p = 4$) are enough to uniquely characterize 90% of individuals. The blue bars represent unicity when using spatial-temporal-price triples ($a = 0.50$) and show that adding the approximate price of a transaction significantly increases the likelihood of reidentification. Error bars denote the 95% confidence interval on the mean.

bakery on 23 September and to the restaurant on 24 September. Searching through the data set reveals that there is one and only one person in the entire data set who went to these two places on these two days. $|S(I_p)|$ is thus equal to 1, Scott is reidentified, and we now know all of his other transactions, such as the fact that he went shopping for shoes and groceries on 23 September, and how much he spent.

Figure 1-6 shows that the unicity of financial traces is high ($\mathcal{E}_4 > 0.9$, green bars). This means that knowing four random spatiotemporal points or tuples is enough to uniquely reidentify 90% of the individuals and to uncover all of their records. Simply anonymized large-scale financial metadata can be easily reidentified via spatiotemporal information.

Furthermore, financial traces contain one additional column that can be used to reidentify an individual: the price of a transaction. A piece of outside information, a spatiotemporal tuple can become a triple: space, time, and the approximate price

of the transaction. The data set contains the exact price of each transaction, but we assume that we only observe an approximation of this price with a precision a we call price resolution. Prices are approximated by bins whose size is increasing; that is, the size of a bin containing low prices is smaller than the size of a bin containing high prices. The size of a bin is a function of the price resolution a and of the median price m of the bin. Although knowing the location of my local coffee shop and the approximate time I was there this morning helps to reidentify me, Fig. 1-6 (blue bars) shows that also knowing the approximate price of my coffee significantly increases the chances of reidentifying me. In fact, adding the approximate price of the transaction increases, on average, the unicity of the data set by 22% (Fig. 1-11, when $a = 0.50$, $\langle \Delta\mathcal{E} \rangle = 0.22$).

The unicity $\mathcal{E}$ of the data set naturally decreases with its resolution. Coarsening the data along any or all of the three dimensions makes reidentification harder. We artificially lower the spatial resolution of our data by aggregating shops in clusters of increasing size v based on their spatial proximity. This means that we do not know the exact shop in which the transaction happened, but only that it happened in this geographical area. We also artificially lower the temporal resolution of the data by increasing the time window h of a transaction from 1 day to up to 15 days. Finally, we increase the size of the bins for price a from 50 to 75%. In practice, this means that the bin in which a \$15.13 transaction falls into will go from \$5 to \$16 ($a = 0.50$) to \$5 to \$34 ($a = 0.75$) (table 1.2).

Figure 1-7 shows that coarsening the data is not enough to protect the privacy of individuals in financial metadata data sets. Although unicity decreases with the resolution of the data, it only decreases slowly along the spatial (v), temporal (h), and price (a) axes. Furthermore, this decrease is easily overcome by collecting a few more points (table1.1). For instance, at a very low resolution of $h = 15$ days, $v = 350$ shops, and an approximate price $a = 0.50$, we have less than a 15% chance of reidentifying an individual knowing four points ($\mathcal{E}_4 < 0.15$). However, if we know 10 points, we now have more than an 80% chance of reidentifying this person ($\mathcal{E}_{10} > 0.8$). This means that even noisy and/or coarse financial data sets along all of the dimensions

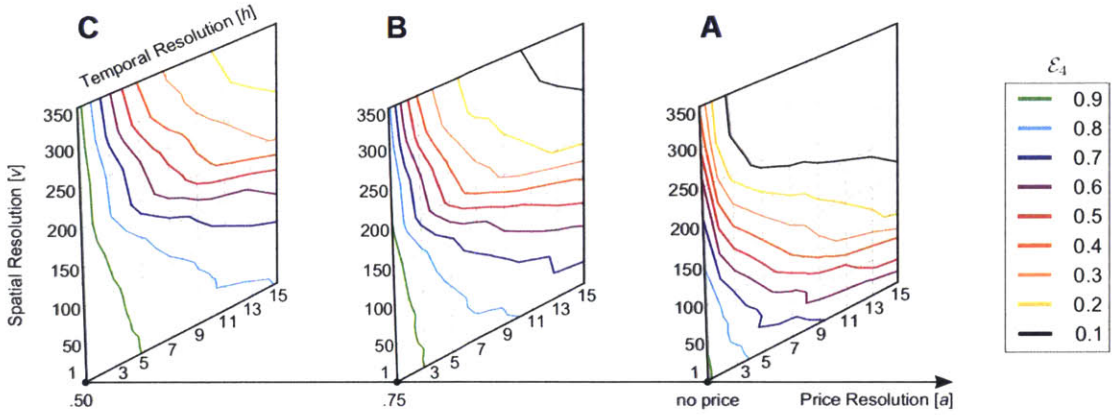


Figure 1-7: Unicity ($\mathcal{E}_4$) when we lower the resolution of the data set on any or all of the three dimensions; with four spatiotemporal tuples [(**A**), no price] and with four spatiotemporal-price triples [(**B**), $a = 0.75$; (**C**), $a = 0.50$]. Although unicity decreases with the resolution of the data, the decrease is easily overcome by collecting a few more points. Even at very low resolution ($h = 15$ days, $v = 350$ shops, price $a = 0.50$), we have more than an 80% chance of reidentifying an individual with 10 points ($\mathcal{E}_{10} > 0.8$) (table 1.1).

provide little anonymity.

We also studied the effects of gender and income on the likelihood of reidentification. Figure 1-8A shows that women are easier to reidentify than men, whereas Fig. 1-8B shows that the higher somebody's income is, the easier it is to reidentify him or her. In fact, in a generalized linear model (GLM), the odds of women being reidentified are 1.214 times greater than for men. Similarly, the odds of high-income people (and, respectively, medium-income people) to be reidentified are 1.746 times (and 1.172 times) greater than for low-income people. Although a full causal analysis or investigation of the determinants of reidentification of individuals is beyond the scope of this paper, we investigate a couple of variables through which gender or income could influence unicity. A linear discriminant analysis shows that the entropy of shops, how one shares his or her time between the shops he or she visits, is the most discriminative factor for both gender and income.

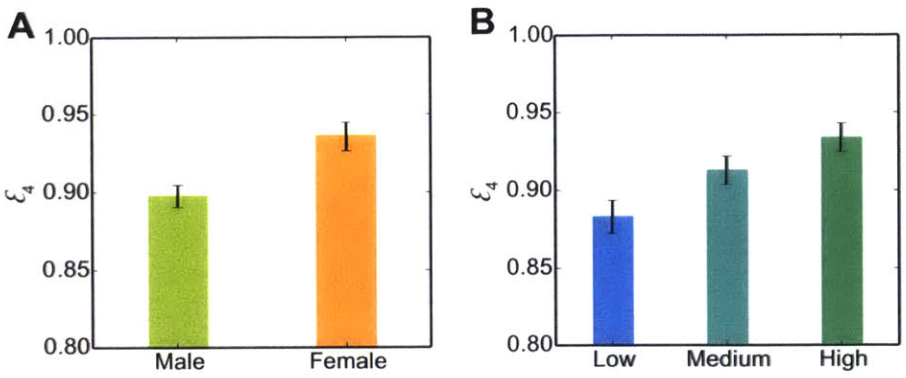


Figure 1-8: Unicity for different categories of users ($v = 1$, $h = 1$). (**A**) It is significantly easier to reidentify women ($\mathcal{E}_4 = 0.93$) than men ($\mathcal{E}_4 = 0.89$). (**B**) The higher a person's income is, the easier he or she is to reidentify. High-income people ($\mathcal{E}_4 = 0.93$) are significantly easier to reidentify than medium-income people ($\mathcal{E}_4 = 0.91$), and medium-income people are themselves significantly easier to reidentify than low-income people ($\mathcal{E}_4 = 0.88$). Significance levels were tested with a one-tailed t test ($P < 0.05$). Error bars denote the 95% confidence interval on the mean.

1.2.3 Discussion

Our estimation of unicity picks the points at random from an individual's financial trace. These points thus follow the financial trace's nonuniform distributions (Fig. 1-9A and Fig. 1-12A). We are thus more likely to pick a point where most of the points are concentrated, which makes them less useful on average. However, even in this case, seven points were enough to reidentify all of the traces considered (Fig. 1-13). More sophisticated reidentification strategies could collect points that would maximize the decrease in unicity.

Although future work is needed, it seems likely that most large-scale metadata data sets—for example, browsing history, financial records, and transportation and mobility data—will have a high unicity. Despite technological and behavioral differences (Fig. 1-9B and Fig. 1-12), we showed credit card records to be as reidentifiable as mobile phone data and their unicity to be robust to coarsening or noise. Like credit card and mobile phone metadata, Web browsing or transportation data sets are generated as side effects of human interaction with technology, are subjected to the same idiosyncrasies of human behavior, and are also sparse and high-dimensional

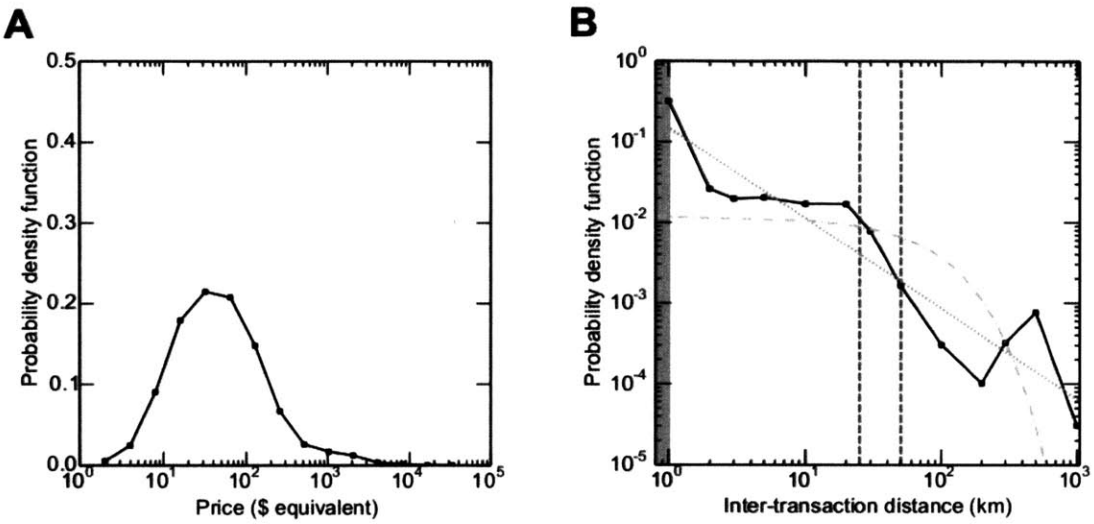


Figure 1-9: Distributions of the financial records. (**A**) Probability density function of the price of a transaction in dollars equivalent. (**B**) Probability density function of spatial distance between two consecutive transactions of the same user. The best fit of a power law (dotted line) and an exponential distribution (dot-dashed line) are given as a reference. The dashed lines are the diameter of the first and second largest cities in the country. Thirty percent of the successive transactions of a user are less than 1 km apart (the shaded area), followed by, an order of magnitude lower, a plateau between 2 and 20 km, roughly the radius of the two largest cities in the country. This shows that financial metadata are different from mobility data: The likelihood of short travel distance is very high and then plateaus, and the overall distribution does not follow a power-law or exponential distribution.

(for example, in the number of Web sites one can visit or the number of possible entry-exit combinations of metro stations). This means that these data can probably be relatively easily reidentified if released in a simply anonymized form and that they can probably not be anonymized by simply coarsening of the data.

Our results render the concept of PII, on which the applicability of U.S. and European Union (EU) privacy laws depend, inadequate for metadata data sets [178]. On the one hand, the U.S. specific-types approach—for which the lack of names, home addresses, phone numbers, or other listed PII is enough to not be subject to privacy laws—is obviously not sufficient to protect the privacy of individuals in high-unicity metadata data sets. On the other hand, open-ended definitions expanding privacy laws to "any information concerning an identified or identifiable person" [73] in the EU proposed data regulation or "[when the] re-identification to a particular person is not possible" [197] for Deutsche Telekom are probably impossible to prove and could very strongly limit any sharing of the data [80].

From a technical perspective, our results emphasize the need to move, when possible, to more advanced and probably interactive individual [85] or group [91] privacy-conscientious technologies, as well as the need for more research in computational privacy. From a policy perspective, our findings highlight the need to reform our data protection mechanisms beyond PII and anonymity and toward a more quantitative assessment of the likelihood of reidentification. Finding the right balance between privacy and utility is absolutely crucial to realizing the great potential of metadata.

1.2.4 Supplementary Materials

Materials and Methods

The dataset. This study was performed on an anonymized financial dataset of credit card transactions for $\sim$ 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 2] 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. 1-6, 1-7 and 1-8, as well as the GLM model and all the unicity statistics.

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 [88] 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. 1-14 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 1 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 1.2 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 $\mathcal{E}_p$ of a dataset by performing a unicity test on t=10,000 sampled users with at least p points, as described in [79]. 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.

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

A pseudo-code for the estimation of the unicity of a dataset is given in Algorithm 2 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

$\langle \Delta\mathcal{E} \rangle$ quantifies how much adding a dimension to the data increases unicity. We compute the average unicity $\langle \Delta\mathcal{E} \rangle$ at different resolutions of space and time over the linearly interpolated surface to avoid effects of sampling. It is interesting to notice in Fig. 1-11 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

	Price Resolution [a]		
	0.50	**0.75**	**no price**
$\mathcal{E}_4$	.13	.06	.00
$\mathcal{E}_6$	.40	.25	.03
$\mathcal{E}_{10}$	.86	.72	.21

Table 1.1: Unicity at very low spatio-temporal resolution ($h = 15$, $v = 350$) knowing four ($\mathcal{E}_4$), six ($\mathcal{E}_6$), and ten ($\mathcal{E}_{10}$) points.

through which gender or income could influence $\mathcal{E}$; 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 1-12 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.

Algorithm 1 Bins(a)

$top \leftarrow .4 + (.4 \cdot a)$
$bins \leftarrow \{.4 - (.4 \cdot a), top\}$
while $top \leq 22800$ **do**
  $bottom \leftarrow top$
  $m \leftarrow bottom/(1 - a)$
  $top \leftarrow m \cdot (1 + a)$
  $bins \leftarrow bins + \{top\}$
end while
return bins

Algorithm 2 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
    end if
  end for
  if is_unique then
    uniqueUsers ← uniqueUsers + {u}
  end if
end for
return |uniqueUsers|/|users|
```

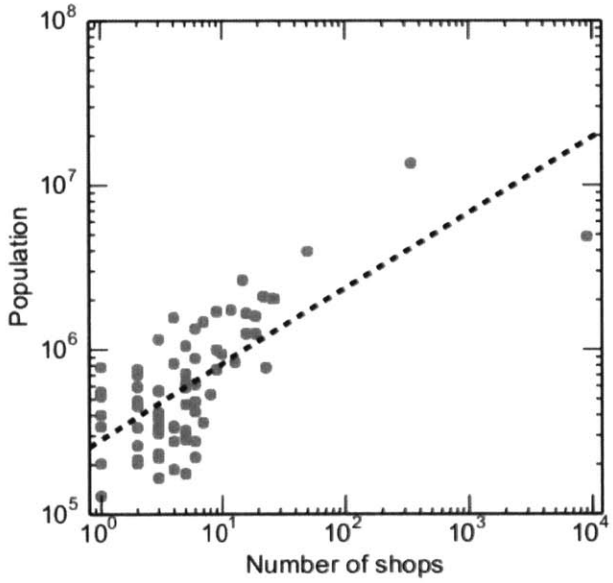


Figure 1-10: The number of shops per district is strongly correlated with its population ($r^2 = 0.51, P < 0.001$). This emphasizes our ability to generalize these results to other financial datasets.

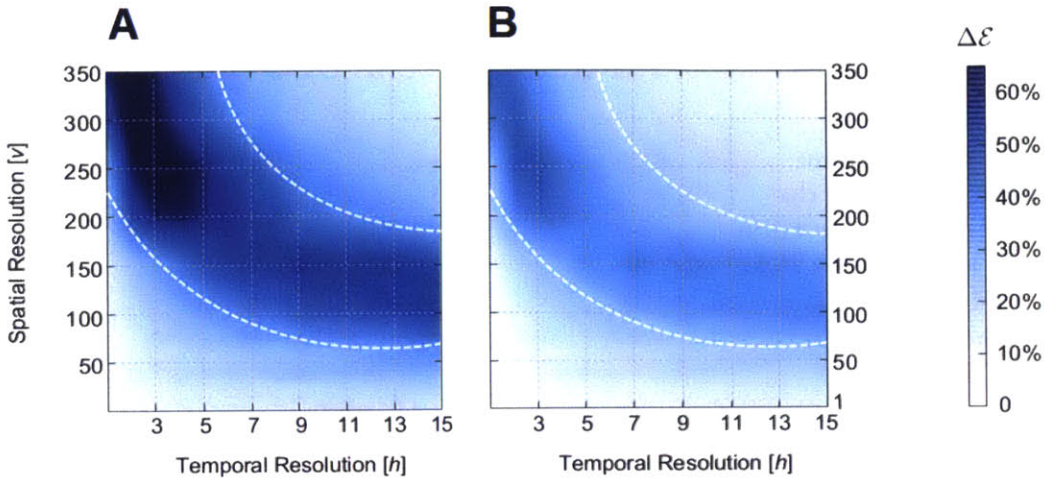


Figure 1-11: Gain in unicity ($\Delta\ \mathcal{E}$) when adding a third dimension, the approximate price of a transaction (**A**, $a = 0.50$; **B**, $a = 0.75$). We see that the gain in unicity $\Delta\ \mathcal{E}$is higher in the central region marked with dashed lines.

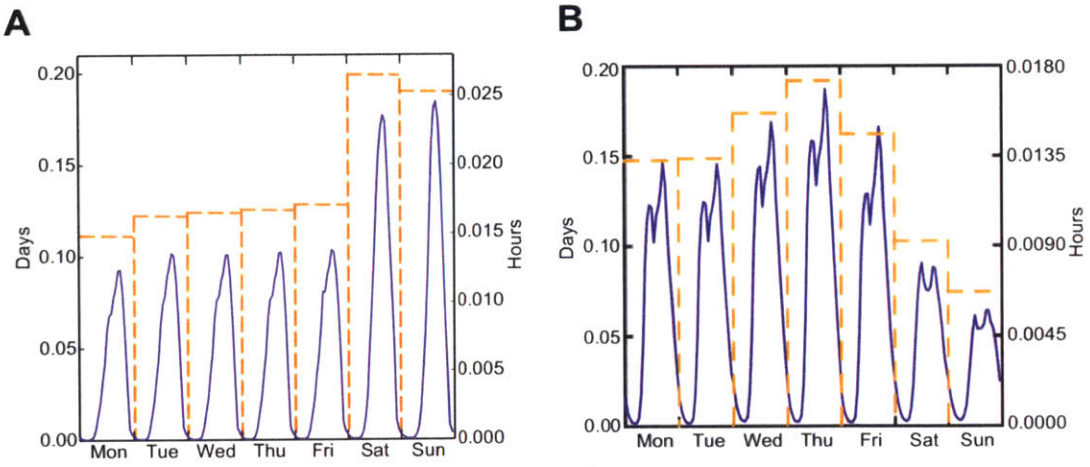


Figure 1-12: (**A**) Probability of having a credit card record per hour (blue right axis) and per day (orange, left axis). (**B**) Probability of having a mobile phone record per hour (blue, right axis) and per day (orange, left axis) as reported in [79].

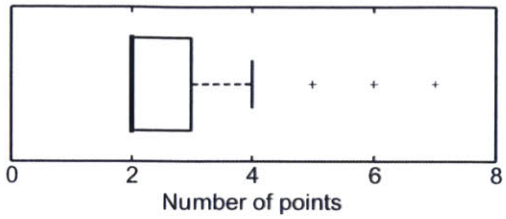


Figure 1-13: While a trace may not be uniquely re-identified with p spatio-temporal-price triples, the same trace might be unique if more triples are known. We here evaluate the minimum number of triples p needed to uniquely characterize every trace in a set of 10,000 randomly sampled traces with at least p points ($h = 1$, $v = 1$, $a = 0.50$). In this set of traces, 7 spatio-temporal-price points are enough to re-identify all of them including the most difficult one.

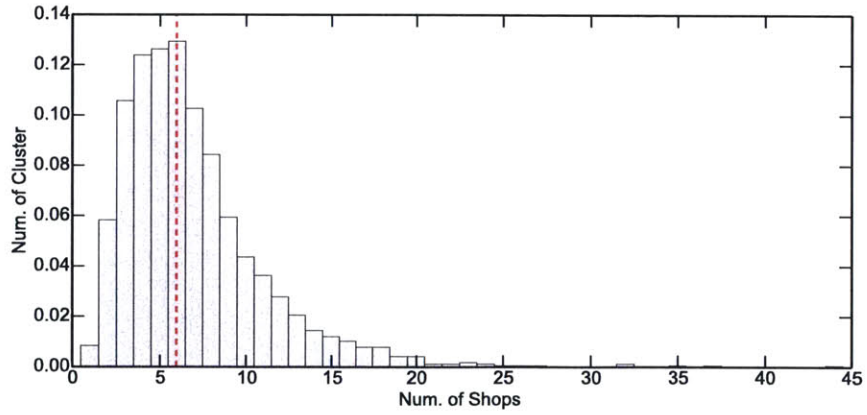


Figure 1-14: Cluster's size resulting from the F.S.C.L. algorithm with $\mu = 6$. The dashed red line indicates the empirical mean of 5.9975.

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 1.2: **(A)** Bins for $a = 0.50$. **(B)** Bins for $a = 0.75$.

Chapter 2

The Risk of Inference

2.1 Personality Prediction from Mobile Phone Data[1]

2.1.1 Introduction

How much can one know about your personality just by looking at the way you use your phone? Determining the personality of a mobile phone user simply through standard carriers' log has became a topic of tremendous interest. Mobile cellular subscriptions have hit 6 billion throughout the world [71] and carriers have increasingly made available phone logs to researchers [94] as well as to their commercial partners [72]. If predicted correctly, mobile phones datasets could thus provide a valuable unobtrusive and cost-effective alternative to survey-based measures of personality. For example, marketing and phone companies might seek to access dispositional information about their customers to design customized offers and advertisements [86]. Appraising users dispositions through automatically collected data could also benefit the field of human-computer interface where personality has become an important factor [38]. Finally, finding ways to extract personality and, more broadly, psychosocial variables from country-scale datasets might lead to unprecedented discoveries in social sciences.

The idea of predicting people's personalities from their cellphone stems from recent advances in data collection, machine learning, and computational social science showing that it is possible to infer various psychological states and traits from the way people use everyday digital technologies. For example, some researchers have shown that pattern in the use of social media such as Facebook or Twitter can be used to predict users' personalities [41, 74, 188]. Others have used information about people's usage of various mobile phone applications (e.g., YouTube, Internet, Calendar, Games, etc.) or social network to draw inferences about phone owners' mood and personality traits [68, 90, 202, 187, 167]. Although these approaches are interesting, they either require to have access to extensive information about people's entire social network or people to install a specific tracking application on their phone. These

[1]Published as de Montjoye, Y.-A.*, Quoidbach J.*, Robic F.*, Pentland A., Predicting people personality using novel mobile phone-based metrics. Proc SBP, Washington, USA (2013)

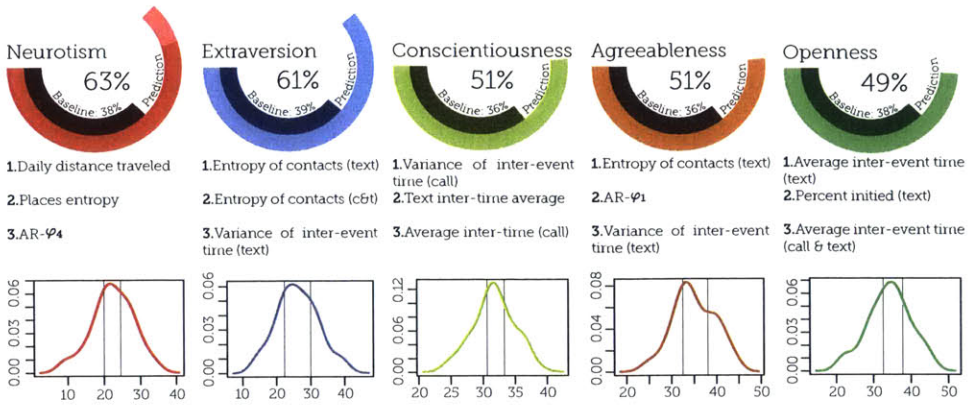


Figure 2-1: **(A)** Accuracy of the prediction with respect to the baseline, **(B)** most useful features to predict personality traits, and **(C)** the distribution of personality traits across our dataset.

constraints greatly undermine the use of such classification methods for large-scale investigations.

The goal of the present research is to show that users' personalities can be reliably inferred from basic information accessible from *all* mobile phones and to *all* service providers. Specifically, we introduce five sets of psychology-informed metrics–Basic phone use, Active user behaviors, Mobility, Regularity, and Diversity–that can be easily extracted from standard phone logs to predict how extroverted, agreeable, conscientious, open to experience, and emotionally stable a user is.

2.1.2 Results

Table 2.1.2 displays the different indicators and their respective contribution in predicting the big 5. Specifically, 36 out of our indicators were significantly related to personality and were all included in the final SVM classifier. As depicted in Figure 2-1, the model predicted whether phone users were low, average, or high in neuroticism, extraversion, conscientiousness, agreeableness, and openness with an accuracy of 54%, 61%, 51%, 51%, and 49%, respectively. The baselines being between 36 and 39%, we predict on average 42% better than random. For neuroticism, the predictive power of the model was further increased by including participants' gender as a predictor,

increasing the accuracy to 63%. This finding is not surprising given that neuroticism is one of the traits that is most strongly associated with gender, with women having higher means levels than men in most countries world-wide [137].

An investigation of the most important feature to predict each trait revealed interesting associations. Indicators linked to users' mobility (i.e., distance traveled and entropy of places) were useful to predict Neuroticism. The entropy of participants' contacts helped predict both Extraversion and Agreeableness. These findings are inline with past research showing these traits both relate to different aspects of the diversity of one's social network: extraverts tend to seek more friends than introverts, agreeable individuals tend to be selected more as friends by other people [180]. Highly consistent with past research showing that conscientious individuals tend to like organization, precision, and punctuality [139], we found that the best predictor of Conscientiousness was the variance of the time between phone calls. Lastly, the strongest predictor of Openness was the average time between text interactions–a finding that remains be explained be future research.

	N	E	O	C	A
Regularity					
Average inter-event time (call)	•	•	•	•	•
Average inter-event time (text)			•	•	
Average inter-event time (c&t)		•	•	•	
Variance of inter-event time (call)		•		•	
Variance of inter-event time (text)		•		•	•
Variance of inter-event time (c&t)	•	•		•	•
Home regularity		•			•
AR- φ_1			•		
AR- φ_4	•				•
AR- φ_8	•	•	•		
AR- φ_{12}				•	•
AR- φ_{24}		•	•	•	
Number of call regularity				•	•
Diversity					
Entropy of contacts (call)			•	•	•
Entropy of contacts (text)	•	•	•	•	•
Entropy of contacts (c&t)		•			
Contacts to interactions ratio (call)	•	•	•		•
Contacts to interactions ratio (text)	•	•		•	
Contacts to interactions ratio (c&t)	•	•	•		
Number of contacts (call)		•	•		
Number of contacts (text)		•		•	
Number of contacts (c&t)		•	•		
Spatial behavior					
Radius of gyration (daily)	•	•			
Distance traveled (daily)	•	•	•	•	•
Number of places	•	•	•	•	•
Entropy (places)	•	•	•	•	•
Active behavior					
Response rate (call)		•			
Response rate (text)	•		•	•	•
Response latency (text)			•		
Percent during the night (call)	•		•		•
Percent initiated (text)	•		•		
Percent initiated (call)	•	•	•	•	
Percent initiated (c&t)				•	
Basic Phone use					
Number of interactions (text)		•			
Number of interactions (call)	•	•	•		•
Number of interactions (c&t)	•	•		•	•

2.1.3 Methodology

Participants and Procedure

The empirical sections of this work are based on a dataset collected from March 2010 to June 2011 in a major US research university [128]. Each participant was equipped with a Android smartphone running the open sensing framework *Funf* [36]. While the framework is designed to collect a wide range of behavioral data from the user's phone, we voluntarily limit ourself to data available in standard carriers's logs such as phone calls, text messages sent and received, etc. These CDR (Call Data Record) have recently become widely use for computational social science research [94, 160, 148, 43, 108]. After removing participants who had less than 300 call or text per year and/or that failed to complete personality measures, our final sample was composed of 69 participants (51% male, Mean age = 30.4, S.D. = 6.1, 1 missing value).

Metrics

We developed a range of novel indicators allowing us to predict users' personality. To build our list of indicators, we examined theories and research in personality psychology and, more specifically, the literature five factor model of personality, the dominant paradigm in personality research [146]. The five-factor model is a hierarchical organization of personality traits in terms of five basic dimensions: Extraversion (i.e, the tendency to seek stimulation in the company of others, to be outgoing and energetic), Agreeableness (i.e, the tendency to be warm, compassionate, and cooperative), Conscientiousness (i.e., the tendency to show self-discipline, be organized, and aim for achievement), Neuroticism (i.e, the tendency to experience unpleasant emotions easily), and Openness (i.e, the tendency to be intellectually curious, creative, and open to feelings).

From this literature review, we generated novel indicators that can be easily computed from carriers logs and that we believed would meaningfully account for potential differences in personality (see Table 2.1.2). These indicators fall under 5 broad categories: Basic phone use (e.g., number of calls, number of texts), Active user behaviors

(e.g., number of call initiated, time to answer a text), Location (radius of gyration, number of places from which calls have been made), Regularity, (e.g.,temporal calling routine, call and text inter-time), and Diversity (call entropy, number of interactions by number of contacts ratio). These indicators are detailed hereafter.

Entropy: Is a quantitative measure reflecting how many different categories there are in a given random variable, and simultaneously takes into account how evenly the basic units are distributed among those categories. For example, the entropy of one's contacts is the ratio between one's total number of contacts and the relative frequency at which one interacts with them. $H(a-c) = -\sum_c f_c \log f_c$ where c is a contact and f_c the frequency at which a communicates with c. The more one interacts equally often with a large number of contacts the higher the entropy will be. This work considers the entropy of calls, text, calls+text but also the entropy of places one visits.

Inter-event time: Is the time elapsed between two events. This work then consider both the average and variance of the inter-event time of ones' call, text, call+text. call+text means that an interaction, a call or an text, happened between two users. Therefore, even though two users have the same inter-event time for both call and text, their mean inter-event times for call+text can be very different.

AR coefficients: We can convert the list of all calls and texts made by a user into a time-series. We discretized time by steps of 6 hours. For example, the time-series X_t contain the number of calls made by a user between 6pm and 12am on Monday followed by the number of calls made by the same user between 12am and 6am on Tuesday and so on. We then train a *auto-regressive* model per user. This model takes the form $X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$ where c is a constant and ε_t are noise terms. The coefficients φ_i can thus be interpreted as the extent to which knowing how many calls a person made in the previous 6 hours, the day before at the same time predicts the number of calls that person will make in the coming 6 hours. We only kept the coefficient that were statistically significant for at least 3 traits: $\varphi_{1,4,8,12,18,24}$. Note that while we see some patterns in the statistically significant coefficients, interpretation of such patterns requires caution given that (1) this analysis has been done post-hoc

and (2) our relatively small sample size.

Response rate and latency (text): We consider a text from a user (A) to be a response to a text received from another user (B) if it is sent within an hour after user A received the last text from user B. The response rate is the percentage of texts people respond to. The latency is the median time it takes people to answer a text. Note than by definition, latency will be less or equal to one hour.

Number of places and their entropy: The dataset was collected using the open sensing framework *Funf* which prevent us from directly using cell phone towers. We instead empirically defined places by grouping together the GPS points of a user that are less than 50m apart and by defining their center of mass as the lat-long coordinate of the place. 50m made sense given the sampling resolution of our dataset. Finally, we only kept the places where a user spend more than 15 minutes in a row.

Radius of gyration: This is the radius of the smallest circle that contains all the places a user have been to on a given day.

Distance per day: This is sum of the distance between the consecutive places a user has visited in a given day.

Home and call regularity: We look at regularity at which a user is coming back home (home regularity) or receiving/making a call (call regularity) using a neural coding inspired metric [211].

Personality

As part of a larger questionnaire, participant completed the Big Five Inventory (BFI-44 [123]), a 44-item self-report instrument scored one a 5-point Likert-type scale measuring the Big Five personality traits. The BFI-44 has been widely used in personality research and has been shown to have excellent psychometric properties [123]. As depicted in Figure 2-1, participants personality scores follow a normal distribution: Neuroticism ($A = 0.3012$, $p = 0.5698$), Openness ($A = 0.2592$, $p = 0.7042$), Extraversion ($A = 0.2884$, $p = 0.6074$), Conscientiousness ($A = 0.4380$, $p = 0.2869$), and Agreeableness ($A = 0.4882$, $p = 0.2162$).

Class prediction

Because the relationship between personality traits and numerous behavioral and psychological factors can often be non-linear [52, 75], we choose to use SVM over the more traditional GLM as the former automatically model non linear relationships. Consequently, following [138] we classified each user as low, average, or high on each on the five personality dimensions.

We then selected the most relevant features using a greedy method similar to [114]. At each iteration, features are ranked using the squared weight and the worst feature of the set is removed. We stop removing features when removing a subset of worst features of size less than 3 degrades the performance and report the 3 highest ranked features. We then classified using an SVM with a 10-fold cross validation.

2.1.4 Discussion

The present study provides the first evidence that personality can be predicted from standard carriers' mobile phone logs. Using a set of novel indicators that we developed based on personality research and that are available to virtually anyone, we were able to predict whether users were low, average or high on each of the big five from 29% to 56% better than random. These levels of accuracy were obtained while we purposefully adopted a restrictive approach only using phone logs.

To our knowledge, these predictions exceed all previous research linking psychological outcomes to mobile phone use. In particular, a previous study that used a combination of information from mobile phone logs *and* people's usage of mobile phone applications such as YouTube, Internet, and Games predicted the personality of their owners with a mean accuracy of 15% [68]. In comparison, the mean accuracy in the present research is almost three times as high (i.e., 42%).

It is interesting to note that Extraversion and Neuroticism were the traits that were best predicted in our study. These two traits are the dimensions of personality that are the most directly associated with emotion. In particular, extraversion is a strong predictor of positive emotions and neuroticism is a strong predictor of negative

emotion [107]. This raises the hypothesis that our indicators might be picking up on the emotional components associated with these two traits. It would be interesting to investigate whether our indicators can predict emotional variable such as happiness in future studies. In addition, contrasting cellphone-based vs. questionnaire-based measures of personality when predicting various psycho-social outcomes might lead to interesting asymmetries. In line with this idea, recent research in personality shows that ratings of one's personality that are made by oneself and ratings of one's personality that are made by others are both valid but different predictors of behavior. For example, self-ratings predict behaviors like arguing or remaining calm, whereas other-ratings predict behaviors like humor and socializing [201].

Although more research is needed to validate our model and the robustness of our indicators for use on a large-scale and more diverse population, we believe that our findings open the door to exciting avenues of research in social sciences. Our personality indicators and the ability to predict personality using readily available mobile phone data may enable cost-effective, questionnaire-free investigation of personality-related questions at the scale of entire countries.

Chapter 3

Privacy-Conscientious Solutions

3.1 Privacy-Conscientious Uses of Mobile Phone Data[1]

Mobile phones are now ubiquitous in developing countries, with 89 active subscriptions per 100 inhabitants [121]. Though many types of population data are scarce in developing countries, the metadata generated by millions of mobile phones and recorded by mobile phone operators can enable unprecedented insights about individuals and societies. Used with appropriate restraint, this data has great potential for good, including immediate use in the fight against Ebola [98].

To operate their networks, mobile phone operators collect call detail records– metadata of who called whom, at what time, and from where. After the removal of names, phone numbers, or other obvious identifiers, this data can be shared with researchers to reconstruct precise country-scale mobility patterns and social graphs. These data have already been used to study importation routes of infectious diseases [209], migration patterns, or economic transactions [93]. Such data are now being actively sought to inform the fight against Ebola [208] but, despite the promise, this effort appears stalled [30, 31].

As part of MIT's Big Data initiative, we examined two operational use cases of mobile phone data for development modeled on previous research. The first case, involved the use of location metadata to understand and quantify the spread of infectious diseases (e.g. malaria or Ebola) within and among countries [30, 31]. The second case considered the use of behavioral indicators derived from mobile phone metadata to micro-target outreach or drive uptake of agricultural technologies or health seeking behavior [191]. Here, mobile phone data could be used to define subgroups based on specific traits and behaviors, which would then receive messages or other outreach from the mobile operator [2]. We also considered cases where the data could be used

[1]Published as de Montjoye, Y.-A., Kendall, J., and Kerry, C. (2014) Enabling humanitarian use of mobile phone data. Brookings, Issues in Technology Innovation, 26.

[2]This is very similar to how some mobile marketing interfaces work where marketers will specify the criteria and identifying characteristics for the people they want to target with specific messages but would not receive actual numbers. Alternatively, anonymized data could be shared with encrypted identifiers which would be passed back to the operator to trigger outreach.

to select individuals to be identified and contacted directly in limited circumstances. These two scenarios are quite distinct from a regulatory and privacy perspective, as we discuss below.

These mobile phone data case studies revealed ways in which, despite the promise, regulatory barriers and privacy challenges are preventing the use of mobile phone metadata from realizing its full potential. More specifically, our analysis showed (1) the lack of commonly-accepted practices for sharing mobile phone data in privacy-conscientious ways and (2) an uncertain and country-specific regulatory landscape for data-sharing especially for cross-border data sharing.

While some forward-looking companies have been sharing limited data with researchers in privacy-conscientious ways, these barriers and challenges are making it unnecessarily hard for carriers to share data for humanitarian purposes [84, 56]. We describe these issues further and offer recommendations moving forward.

3.1.1 Protecting the Identity of Subjects

Mobile phone metadata made available to researchers should never include names, home addresses, phone numbers, or other obvious identifiers. Indeed, many regulations and data sharing agreements rely heavily on protecting anonymity by focusing on a predefined list of personally-identifiable information that should not be shared. In the United States, for example, the privacy rule issued by the Department of Health and Human Services to protect the privacy of patient health records specifies 18 different types of data about patients that must be removed from datasets for them to be considered de-identified [159].

However, elimination of specific identifiers is not enough to prevent re-identification. The anonymity of such datasets has been compromised before and research [79] shows that, in mobile phone datasets, knowing as few as four data points–approximate places and times where an individual was when they made a call or send a text–is enough to re-identify 95% of people in a given dataset. In general, there will be very few people who are in the same place at the same time on four different occasions, which creates a unique "signature" for the individual making it easy to isolate them as unique in

the dataset. The same research also used unicity to shows that simply anonymized mobile phone datasets provide little anonymity even when coarsened or noised.

This means that removing identifying information makes isolating and identifying a specific person in the dataset only slightly more challenging because that person can be identified using available sources of data that link location with a name or another identifier (e.g. geo-tagged posts on social media, travel schedules, etc.). Wholesale re-identification is more difficult, however, because re-identification of a large fraction of the dataset requires access to a full list of people and places they have been, which may not be as easy to acquire. Nevertheless, a determined attacker can still re-identify people using such data. Therefore, removing personally identifiable information is only a first step in most instances and more stringent approaches are required unless trust in the recipient of a dataset is high.

Recognizing the limits of an approach to anonymity and re-identification that focuses only on identity information like names or ID numbers, governments have sought to expand protection beyond identity to any information that can be used to identify an individual. In 2007, the federal Office of Management and Budget added to its list of identifiers "any other personal information which is linked or linkable to an individual." [99] In Europe, the Directive 95/46/EC cautions that "account should be taken of all the means likely to be used" to identify an individual, [200] and a thorough recent opinion of EU privacy regulators provided technical guidance on the challenges and risks of re-identification [165].

The challenge of these broad definitions is that they are open-ended. No existing anonymization methods or protocols can guarantee at 100 percent that mobile phone metadata cannot be re-identified unless the data has been greatly modified or aggregated. Hence, open-ended requirements can be unverifiable and, taken to their logical extreme, so strict as to prohibit any sharing of data even when risk of re-identification is very limited.

We believe this places too much emphasis on a limited risk of re-identification and unclear harm without considering the social benefits of using this data such as better managing outbreaks or informing government response after a disaster [51]. Special

consideration should be given to cases where the data will be used for significant public good or to avoid serious harm to people. Furthermore, data sharing should allow for greater levels of disclosure to highly trusted data recipients with strong processes, data security, audit, and access control mechanisms in place. For example, trusted third parties at research universities might warrant access to richer, less anonymized data for research purposes and be relied on not to try to re-identify individuals or to use the data inappropriately.

For both use cases, we defined data-sharing protocols that would allow for the intended analysis, while protecting privacy. We contemplate releasing anonymized data to research teams and NGOs in a form that adds technical difficulty to re-identification, limits the amount of data that would be re-identified, and further limiting the risk of re-identification or abuse with a legal agreement that specifies that only specific purposes and other protocols can be applied to the data. In our analysis, we focused on a middle ground scenario of relatively open sharing of data with multiple research teams and/or NGOs, with some (but limited) accountability and auditability. We did not consider a fully-public release where a very high level of anonymization would be required, nor a release to a highly trusted third party with strong data protection in place that might allow weakly-anonymized data sharing.

For our first use case, we concluded that a 5 percent sampling of the data on a monthly basis, resampled with new identifiers every month for a year and coarsened temporally and spatially into 12-hour periods (7 a.m. to 7 p.m.) and by regions within countries would be the right balance between utility and privacy [3]. It would adequately show individuals' mobility across regions under study and the number of nights spent in infected regions while providing significant–but not absolute–protection of identity and limiting the amount of data that would be re-identified.

[3]The back-of-the envelope reasoning goes as: We use a spatial resolution of 17 antennas on average ($v = 17$) and a temporal resolution of 12 hours ($h = 12$). This means that with 4 points in a given month, we'd have a approx. 20% chance ($\mathcal{E} = .20$) at re-identifying an individual in a given month (resp. $\mathcal{E} = .55$ with 10 points)(see `http://www.nature.com/srep/2013/130325/srep01376/fig_tab/srep01376_F4.html`). This means that, to have between 20% to 55% chances of re-identifying an individual, we'd need 4 to 10 points every month meaning 48 to 120 points total for a year. Even in this case, as we use a 5% sampling and we resample every month, an individual has only a 45% chance to be in at least one of the sampled month ($1 - 0.95^{12 months}$).

For our second use case, we concluded that the behavioral indicators [77] derived from metadata can be shared with the researchers safely, provided outliers have been removed. Researchers could then use this data to segment the population into specific sub-groups based on traits like calling patterns, mobility, number of contacts, etc. People fitting these criteria could then be contacted by the mobile phone operators through text messages or other communications. Their phone numbers would be known only to the mobile phone operators.

We also considered cases where specific individuals could be contacted based on criteria applied to the data. To do so would require either (a) including in the dataset pseudonymous–but unique–identifiers that make it possible to connect data showing certain traits (such as a likely exposure to disease based on travel patterns) with specific individuals, or (b) including telephone numbers in the dataset so that researchers and/or NGOs can contact the individuals identified directly. Because it enables re-identification, the former would be a departure from good privacy practices unless the data recipient were highly trusted, and the second would be a clear departure because it disclosed unmodified personally identifiable information.

Nevertheless, re-identification could be vital in case of emergencies such as an earthquake [51]. These alternate use cases illustrate further the need to develop mechanisms for trusted third parties to maintain data under strong controls for use, access, security, and accountability [4].

More generally, promising computational privacy approaches to make the re-identification of mobile phone metadata harder include sampling the data, making the antenna GPS coordinates less precise through voronoi translation for example [76], or limiting the longitudinality of the data to cover shorter periods of time. These could go as far as to set up systems or collaborations where researchers could pose questions of the data, but where mobile operators would only share with researchers "answers," [78, 85] such as behavioral indicators or summary statistics [5]. Each of

[4]We assume here that the mobile operator does not have explicit permission from the data subject to disclose their information. If users were to opt-in to sharing this would then become permissible.

[5]While promising, these solutions are not yet ready for prime-time. Standardized software to process call detail records along with testing and reporting tools are still under development while the use of online systems allowing researchers to ask questions that would be run against the data

these alternatives could be employed depending on the use the data is put to, the amount and sensitivity of the data that would be uncovered, how and by whom the data will be governed and housed, and the attendant risks of harm.

3.1.2 Engaging Government Support

The second challenge we identified to humanitarian use of mobile phone metadata is an uncertain and country-specific regulatory landscape for data-sharing. Our study focused on Africa, where data privacy regulation has been evolving along two lines. The Francophone countries–mostly located in West Africa, where current exposure to Ebola is greatest–have tended to adopt privacy frameworks modeled on the 1995 European Privacy Directive and supervised by national data protection authorities. Meanwhile English-speaking countries with common law systems either have not yet adopted comprehensive privacy laws, or have adopted country-specific laws.

This landscape presents a number of barriers to humanitarian use of mobile phone metadata. First, legal uncertainty complicates the design of data-sharing protocols. Indeed, even in countries that have had laws and regulatory agencies in place for some time, the relevant rules have not developed in enough detail to address an issue that is often uncertain even in the most developed legal systems.

Second, as discussed above, questions about the validity of most methods of de-identification persist particularly in countries that use open-ended definitions of anonymization such as the EU one. There exist no widely accepted data-sharing standards to help various actors achieve a rational privacy/utility tradeoff in using mobile phone metadata.

Third, regardless of legal systems, compatible data-sharing protocols–including data de-identification–have to be designed and validated on a country by country basis. For example, data-sharing protocols have to be compatible, which includes having both the phone number and the mobile phone identifier [6] hashed with the

and only receive answers would imply architectures investments from mobile phone operators

[6]IMEI or International Mobile Station Equipment, a unique number that identifies a mobile phone on the network.

same function and salt [7] to allow for mobile phones to be followed across border, even if the user changes SIM cards. These issues make cross-border data sharing or intra-regional tracking of population flows particularly complex and costly. Yet such cross-country sharing is essential in the fight against diseases such as malaria or the current Ebola outbreak [157].

Fourth, our second use case contemplated that, in general, only behavioral indicators derived from carriers' metadata would be shared with researchers but that, in specific and limited circumstances where these indicators show an individual would benefit from intervention, the identity could be used to enable remote intervention such as targeted texts sent by the operator, or identification through mechanisms that carefully control the release and use of this information.

In the absence of explicit consent from users to such disclosure and use of data from their mobile phones, these forms of re-identification of data subjects presents obvious privacy challenges and may come into conflict with most privacy legal regimes absent specific exceptions. The EU Privacy Directive provides that data processing must have a lawful basis, but that such a basis may be "to protect the vital interests of the data subject," or "in the public interest, or in the exercise of official authority, and recognizes "public health" as such a public interest." [8] Thus, it will take the support of national governments, their health ministries, and their data protection authorities to enable use of data especially in such exigent situations, but also for a range of humanitarian applications [9].

[7] One potentially interesting solution here would be to rely on multiple hash functions that can be nested.

[8] European Union, Directive 95/46/EC, Article 7 (d), (e). An update to this legislation, the Privacy Regulation proposed by the European Commission in 2012, `http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf`, also included an exception from certain requirements for "scientific, historical, statistical, and scientific research purposes," but this was removed from legislation as passed by the European Parliament. `http://www.europarl.europa.eu/meetdocs/2009_2014/documents/libe/pr/922/922387/922387en.pdf`.

[9] Under the World Health Organization's International Health Regulations, the WHO and member states undertake to conduct "surveillance" for public health purposes and member states are permitted to "disclose and process personal data where essential for purposes of assessing and managing public health risks." WHO, Fifty-eighth World Health Assembly Resolution WHA58.3: Revision of the International Health Regulations, Articles 1 (definition of surveillance), 5.4, and 45 . 2005, `http://www.who.int/ipcs/publications/wha/ihr_resolution.pdf`.

3.1.3 Conclusion: Roadmaps Needed

These privacy challenges and regulatory barriers are making humanitarian data-sharing much harder than it should be for mobile phone operators and are significantly limiting greater use of mobile phone metadata in development or aid programs and in research areas like computational social science, development economics, and public health.

To realize the potential of this data for social good, we recommend the following:

1) There is a clear need for companies, NGOs, researchers, privacy experts, and governments to agree on a set of best practices for new privacy-conscientious metadata sharing models in different development use cases–a wider and higher-level discussion of the kind our MIT working group conducted. These best practices would help carriers and policymakers strike the right balance between privacy and utility in the use of metadata and could be instantiated by data-protection agencies, institutional review boards, and in data protection laws and policies. This would make it easier and less risky for carriers to support humanitarian and research uses of this data, and for researchers and NGOs to use these metadata appropriately.

2) Such best practices should accept that there are no perfect ways to de-identify data–and probably will never be. [154] There will always be some risk that must be balanced against the public good that can be achieved. While much more research is needed in computational privacy, widespread adoption of existing techniques as standards could enable this trend of sharing data in a privacy-conscientious way.

3) Standards and practices as well as legal regulation also need to address and incorporate trust mechanisms for humanitarian sharing of data in a more nuanced way. Protection of individual privacy includes not only protection against re-identification, but also data security and protection against unwanted uses of data. Risk of re-identification is not a purely theoretical concept nor is it binary and it should be assessed vis-à-vis the level of trust placed in the data recipient and the strength of their systems and processes. Tracking of migration patterns or analysis of behavior patterns may offer enormous benefits for disease prevention and treatment, but it is

possible to envision more malignant uses by actors ranging from disgruntled employees of the data recipient to authoritarian governments. The recognition of trusted third-parties and systems to manage datasets, enable detailed audits, and control the use of data could enable greater sharing of these data among multiple parties while providing a barrier against risks.

There is a need for governments to focus on adopting laws and rules that simplify the collection and use of mobile phone metadata for research and public good purposes. Governments should also seek to harmonize laws on the sharing of metadata with common identifiers across national borders. The African Union took what could be a step in this direction last June, when it approved the African Convention on Cyber Security and Personal Data Protection seeking to advance Africa's digital agenda and harmonize rules among African nations [9]. The treaty, which will not take effect until adopted by 15 member states, commits members to adopting a legal framework that follows the template of the European Privacy Directive. Clear and consistent rules will help but only provided they take a pragmatic and privacy-conscientious approach to anonymization, cross-border transfers, and novel uses that enable public good uses of data and allow for public health emergencies and other valuable research.

Research based on mobile phone data, computational privacy, and data protection rules all may seem secondary when confronted by the challenges of poverty, disease, and basic economic growth. But they are on the critical path to realizing the great potential of information technology to help address these critical problems.

3.2 Privacy-Conscientious Data Release: D4D-Senegal[10]

3.2.1 Introduction

There are Big Hopes associated with Big Data: it has been dubbed the oil of the digital economy [166], the next big thing in medical care [189], and a vital tool for building smart cities [18]. In science, the availability of large-scale behavioral datasets has even been compared to the invention of the microscope [129].

There is little doubt that impressive work has already been done by the computational social science and mobile phone research communities. Metadata has, for example, been used to better understand the propagation of malaria, to monitor poverty [209, 94], to analyse human mobility [108], and to study the structure of social communities at a national level [183]. Big Data has, however, to be made more broadly available to further realize its promises. Understanding context remains critical, particularly for a sound interpretation and solution of practical questions. Development economists, urban planners, sociologists, and NGOs need to become familiar with this data. "Inanimate data can never speak for themselves, and we always bring to bear some conceptual framework, either intuitive and ill-formed, or tightly and formally structured, to the task of investigation, analysis, and interpretation" [109].

This is why, in 2012, Orange launched the Data For Development challenge in partnership with the University of Louvain and MIT. D4D-Cote d'Ivoire made five months of mobile phone metadata available [56]. The results were impressive: 260 applications from around the world were submitted to access the data and, after three months, more than 80 research papers had been produced [8]. These papers covered topics as diverse as optimizing bus routes, analyzing social divisions [54], and studying disease containment policies [135].

We are now launching, in collaboration with Sonatel Senegal, the second challenge: **D4D-Senegal** where selected teams will have access to one year of metadata for up to 300,000 people across Senegal. This paper describes the data pre-processing and the

[10]Published as de Montjoye, Y. A., Smoreda, Z., Trinquart, R., Ziemlicki, C., and Blondel, V. D. (2014). D4D-Senegal: the second mobile phone data for development challenge. arXiv preprint arXiv:1407.4885.

three datasets that will be made available, as well as a set of research questions that have been suggested by local partner organizations. More details and the application to participate in the challenge are available at `http://www.d4d.orange.com`.

3.2.2 Data preprocessing

The Call Detail Records (CDR) have been collected for a year, from January 1 to December 31, 2013. The customer identifiers were anonymized by Sonatel before the data was transferred to Orange Labs who did the preprocessing.

The original dataset contained more than 9 million unique aliased mobile phone numbers. When preparing datasets, we retained only users meeting both of these criteria:

1. users having more than 75% days with interactions per given period (biweekly for the second dataset, yearly for the third dataset)

2. users having had an average of less than 1000 interactions per week. The users with more than 1000 interactions per week were presumed to be machines or shared phones.

For commercial and privacy reasons, we do not release the real geographical coordinates of the site where BTSs, the mobile network antennas, are located. Note that several BTS can be co-located. We assigned a new position to each site uniformly in its Voronoi cell (the region consisting of all points closer to that antenna than to any other) to make it harder to re-identify users [76]. The `SITE_ARR_LATLON.csv` file contains the new, noisy, latitude and longitude of the site.

For example:

```
site_id,arr_id,lon,lat
1,2,-17.5251,14.74683
2,2,-17.5244,14.74743
3,2,-17.5226,14.7452
4,2,-17.5164,14.74673
```

3.2.3 Datasets

Simply anonymized mobile phone datasets have been shown to be re-identifiable. For instance, it is possible to find a user in a large-scale mobility data using only four spatio-temporal points and coarsening the data only makes it slightly harder [79].

To balance the potential of the data being broadly used with the risks of re-identification we provide three sampled and aggregated datasets for this challenge:

- **Dataset 1:** One year of site-to-site traffic for 1666 sites on an hourly basis,
- **Dataset 2:** Fine-grained mobility data (site level) on a rolling 2-week basis with bandicoot behavioral indicators at individual level for about 300,000 randomly sampled users meeting the two criteria mentioned before for each 2 week period,
- **Dataset 3:** One year of coarse-grained (123 arrondissement level) mobility data with bandicoot behavioral indicators at individual level for about 150,000 randomly sampled users meeting the two criteria mentioned before for a year,

Each dataset has been designed to balance utility with privacy, utility beeing the research that can be done with the data while privacy is the potential risk of re-identification of users. Datasets are thus either precise spatially and temporally but limited in the time they span (dataset 2), or aggregated geographically (dataset 3) or across users (dataset 1) but covering a longer period of time. Finally, precomputed indicators are provided to help inform behavioral research. Columns that might help re-identification have been 3-anonymized when binned to remove outliers [193].

Note that a fourth dataset of synthetic data will be made available in September and will be described in a future paper.

Individual indicators

Mobility datasets 2 and 3 are supplemented with behavioral indicators from [82] computed from metadata using the bandicoot toolbox [77].

The indicators we provide are:

- `active_days_callandtext_mean`

- active_days_callandtext_sem
- duration_of_calls_mean_mean
- duration_of_calls_mean_sem
- entropy_of_contacts_call_mean
- entropy_of_contacts_call_sem
- entropy_of_contacts_text_mean
- entropy_of_contacts_text_sem
- entropy_of_contacts_callandtext_mean
- entropy_of_contacts_callandtext_sem
- entropy_places_callandtext_mean
- entropy_places_callandtext_sem
- interactions_per_contact_callandtext_mean_mean
- interactions_per_contact_callandtext_mean_sem
- interactions_per_contact_call_mean_mean
- interactions_per_contact_call_mean_sem
- interevents_callandtext_mean_mean
- interevents_callandtext_mean_sem
- interevents_call_mean_mean
- interevents_call_mean_sem
- interevents_text_mean_mean
- interevents_text_mean_sem

Places are in this case sites and nocturnal is defined as 7pm to 7am. A full description of the indicators can be found on the bandicoot document in the data repository and the indicator files have been 3-anonymized on binned data on specific columns to remove outliers [193].

Dataset 1: Antenna-to-antenna traffic

This dataset contains the traffic between each site for a year.

The files `SET1V_M01.csv` through `SET1V_M12.csv` contain monthly voice traffic between sites and are structured as follow:

- **timestamp:** day and hour considered in format YYYY-MM-DD HH (24 hours format)
- **outgoing_site_id:** id of site the call originated from
- **incoming_site_id:** id of site receiving the call
- **number_of_calls:** the total number of calls between these two sites during this hour
- **total_call_duration:** the total duration of all calls between these two sites during this hour

For example:

```
timestamp, outgoing_site_id, incoming_site_id,...
...number_of_calls, total_call_duration
2013-04-01 00,2,2,7,138
2013-04-01 00,2,3,4,136
2013-04-01 00,2,4,7,121
2013-04-01 00,2,5,13,272
2013-04-30 23,1651,1632,1,3601
2013-04-30 23,1653,575,1,20
2013-04-30 23,1653,1653,2,385
```

```
2013-04-30 23,1659,608,1,3601
```

The files `SET1S_M01.csv` through `SET1S_M12.csv` contain monthly text traffic between sites and are structured as follow:

- **timestamp:** day and hour considered in format YYYY-MM-DD HH (24 hours format)
- **outgoing_site_id:** id of site the text originated from
- **incoming_site_id:** id of site receiving the text
- **number_of_sms:** the total number of texts between these two sites during this hour

For example:

```
timestamp, outgoing_site_id, incoming_site_id, number_of_sms
2013-05-01 00,2,12,6
2013-05-01 00,2,14,1
2013-05-01 00,2,21,1
2013-05-01 00,2,28,9
2013-05-31 23,1653,190,2
2013-05-31 23,1653,314,3
2013-05-31 23,1653,367,8
2013-05-31 23,1653,520,1
2013-05-31 23,1653,558,2
```

Note that calls spanning multiple time slots are considered to be in the time slot they started in and only calls or texts between Sonatel customers are taken into account.

The latitude and longitude of the sites is provided in `SITE_ARR_LATLON.csv`.

Dataset 2: Fine-grained mobility

This second dataset contains the trajectories at site level of about 300,000 randomly selected users meeting the two criteria mentioned before over two-week periods. The site locations are provided in `SITE_ARR_LATLON.csv`.

The files `SET2_P01.csv` through `SET2_P25.csv` contain the `user_id`, `timestamp`, and `site_id` for each of the 25 two-week periods. The second digits of the minutes and all the seconds of the timestamps have been replaced with zeros (format YYYY-MM-DD HH:M0:00) For each period, a new sample of about 300,000 users was selected and their `user_id` scrambled. Note that this mean that even if a user were to appear in two periods, he would have a different id, and vice versa, the same id in two periods does not mean that it is the same person.

For example:

```
user_id,timestamp,site_id
1,2013-03-18 21:30:00,716
1,2013-03-18 21:40:00,718
1,2013-03-19 20:40:00,716
1,2013-03-19 20:40:00,716
1,2013-03-19 20:40:00,716
1,2013-03-19 20:40:00,716
1,2013-03-19 21:00:00,716
1,2013-03-19 21:30:00,718
1,2013-03-20 09:10:00,705
1,2013-03-21 13:00:00,705
```

The indicators are computed, for every user, over the course of the two week, and are available in the files `INDICATORS_SET2_P01.csv` through `INDICATORS_SET2_P25.csv`.

Dataset 3: Coarse-grained mobility

This third dataset contains the trajectories at arrondissement level of 146,352 randomly selected users meeting the two criteria mentioned before on a yearly basis.

```
user_id,timestamp,arrondissement_id
37509,2013-01-29 15:00:00,3
84009,2013-01-14 07:00:00,3
84009,2013-01-14 07:00:00,3
84009,2013-01-14 07:00:00,3
80150,2013-01-27 16:50:00,3
52339,2013-01-09 19:50:00,48
52339,2013-01-06 17:50:00,48
52339,2013-01-13 15:40:00,48
52339,2013-01-03 19:00:00,48
52339,2013-01-07 01:30:00,48
```

The files `SET3_M01.csv` through `SET3_M12.csv` contain the `user_id`, `timestamp`, and `arrondissement_id` month by month. The second digits of the minutes and all the seconds of the timestamps have been replaced with zeros (format YYYY-MM-DD HH:M0:00) The indicators are computed, for every user, on a monthly basis. They are available in the files `INDICATORS_SET3_M01.csv` through `INDICATORS_SET3_M12.csv`.

The arrondissement shapefile is provided (`SHAPEFILE_SENEGAL.zip`) as well as a summary table (`SENEGAL_ARR.csv`).

The summary table contains:

- **ARR_ID:** the arrondissement_id
- **REG:** the name of the region
- **DEPT:** the name of the department
- **ARR:** the name of the arrondissement

For example:

```
ARR_ID,REG,DEPT,ARR
1,DAKAR,DAKAR,PARCELLES ASSAINIES
2,DAKAR,DAKAR,ALMADIES
```

3,DAKAR,DAKAR,GRAND DAKAR

4,DAKAR,DAKAR,DAKAR PLATEAU

5,DAKAR,GUEDIAWAYE,GUEDIAWAYE

6,DAKAR,PIKINE,PIKINE DAGOUDANE

Contextual data

- GIS shapefiles for Senegal: Administrative divisions of Senegal shapefiles provided by the ADSN are included in the data package `SHAPEFILE_SENEGAL.zip`
- Weather data: `http://www.wunderground.com/weather-forecast/Senegal.html`
- Demographic and socio-economic data: `http://donnees.ansd.sn/en/BulkDownload`
- Import/Export data: `http://atlas.media.mit.edu/explore/tree_map/hs/export/sen/all/show/2010/`
- More references at: `http://www.d4d.orange.com/en/partners-resources/resources`

3.3 Privacy-through-Security: On the Trusted Use of Data[11]

3.3.1 Motivation

Personal Data has become the new oil of the Internet [23], and the current excitement about Big Data is increasingly about the analysis of personal data: location data, purchasing data, telephone call patterns, email patterns, and the social graphs of LinkedIn, Facebook, and Yammer. However, currently personal data is mostly siloed within large companies. This prevents its use by innovative services and even by the user who generated the data. The problem is that while there is substantial legal and social policy scholarship concerning ownership and fair use of personal data, a pragmatic technical solution that allows governments and companies easy access to such data and yet protects individual rights and privacy has yet to be realized and tested.

We therefore develop and test an architecture for the trusted use of large-scale personal data that is consistent with new "best practice" standards which require that individuals retain the legal rights of possession, use, and disposal for data that is about them. To accomplish this, we develop openPDS–an open-source Personal Data Store enabling the user to collect, store, and give access to their data while protecting their privacy. Via an innovative framework for third-party applications to be installed, the system ensures that most processing of sensitive personal data takes place within the user's PDS, as opposed to a third-party server. The framework also allows for PDSs to engage in privacy-preserving group computation, which can be used as a replacement for centralized aggregation.

Although our aim is to provide a technical solution, it is important for such solution to be not only compatible but also aligned with political and legal thinking. openPDS is compatible with and incorporates best practice suggestions of the US Consumer Privacy Bill of Rights [27], the US National Strategy for Trust Identities

[11]Published as de Montjoye, Y. A., Wang, S. S., Pentland, A. (2012). On the Trusted Use of Large-Scale Personal Data. IEEE Data Eng. Bull., 35(4), 5-8.

in Cyberspace (NSTIC) [24], the Department of Commerce Green Paper, and the Office of the Presidents International Strategy for Cyberspace [21]. In addition, it follows the Fair Information Practices (FIPs) which have mandated that personal data be made available to individuals upon request. In addition openPDS is aligned with the European Commission's 2012 reform of the data protection rules [16]. This reform redefines personal data as "any information relating to an individual, whether it relates to his or her private, professional or public life." It also states the right for people to "have easier access to their own data and be able to transfer personal data from one service provider to another more easily" as well as a right to be forgotten. All these ideas and regulations recognize that personal data needs to be under the control of the user in order to avoid a retreat into secrecy where these data become the exclusive domain of private companies, denying control to the user.

3.3.2 Personal Data Stores (PDS)

Many of the initial and critical steps towards implementation of these data ownership policies are technological. The user needs to have control of a secured digital space, a personal data store (PDS), where his data can live. Given the huge number of sources of data that a user interacts with every day, mere interoperability is not enough. There needs to be a centralized location that a user is able to view and reason about the data that is collected about himself. The PDS should allow the user to easily control the flow of data and manage fined grained authorizations for third-service services, fulfilling the vision of the New Deal on Data [23]. A PDS-based market is likely to be fair, as defined by the Fair Information Principles, as the user is the one controlling the access to his data. The user can decide whether such services provide enough value compared to the amount of data it asks for; the user can ask questions like "Is finding out the name of this song worth enough to me to give away my location?" The PDS will help the user make the best decision for himself. Using a privacy-preserving PDS allows for greater data portability, as the user can seamlessly interface new services with his PDS, and will not lose ownership or control of his personal data.

Thanks to the policy requirement of data portability, a PDS-based data market is likely to be economically efficient, as the system removes barriers to entry for new businesses. It allows the more innovative companies to provide better data-powered services. The services chosen by the user will have access to historical data, which was potentially collected even before the creation of the service. Moreover, the services will not be forced to collect data themselves, as they will have access to data coming from other apps. Service providers can thus concentrate on delivering the best possible experience to the user. For example, a music service could provide you a personalized radio station, leveraging the songs and artists you said you like across the web, what your friends like, or even which nightclubs you go to. The real value of large-scale data appears when innovators can create data driven applications on top of rich personal user information.

3.3.3 Question Answering Framework

In the existing mobile space, personal data is offloaded from mobile devices onto servers owned by the application creator. This model prevents users from being able to control their own data; once they hand that data over to a corporation, it is difficult or impossible to refute or retract.

The key innovation in the openPDS model is that computations on user data are performed in the safe environment of the PDS, under the control of the user. The idea is that only the relevant summarized data for providing functionality to the application should leave the boundaries of the user's PDS.

Rather than exporting raw GPS coordinates, it could be sufficient for an app to know which general geographic zone you are currently in. Instead of sending raw GPS coordinates to the app owner's server to process, that computation can be done by the PDS app in the user's PDS server. The system is still exposing personal data of the user, but it is constrained to be what the app strictly needs to know, rather than the raw data objects the user generates. A series of such computed answer would also be easier to anonymized than high-dimensional sensor data. App designers would take care to declare to users as well as in a machine readable format to be enforced exactly

what data is being computed over, what inferences are being exposed to external apps, and what data is being reported back to the company's servers.

With this model of computation, it is relatively easy to monitor the communication between a PDS app and its Android counterpart. Since the user owns the platform on which the PDS app executes, it is possible to eavesdrop on the data that is exposed by the PDS app to the Android app. If an app is accessing and exporting more data than it is supposed to be in order to provide the required services, it will be known by people who use the app, and could potentially be reflected in the app's reviews. This ability to monitor the results of computation on user data provides a coarse way to verify that one's personal data is not being unexpectedly leaked.

3.3.4 The user experience

If Alice chooses to download a PDS-aware version of Spotify, the music streaming service, she would install it just like she would any other Android application. Upon launching the application, the Android app would prompt her to install a Spotify app onto her PDS. The description of the PDS app would describe exactly what data Spotify would access and reason over on her PDS, as well as what relevant summarized information is passed on to Spotify's servers, for example to offer personalized music radios to the user. This allows Alice to understand what it means for her privacy to install the app.

When using the Spotify Android app, rather than storing Alice's personal data on Spotify's servers, the Spotify PDS app would instead access and process the data on Alice's PDS. Alice would have installed or bought a PDS instance on her favorite cloud provider, or on her own server. Over time, her PDS would be filled with information collected by her phone, but also information about her musical tastes, her contacts, as well as a stream of other sensor information that Alice accumulates in her day to day life. Alice would have full control over this data, and could see exactly what data her phone, other sensors, and services gathers about her over time.

Because the Spotify PDS app is being run on a computing infrastructure that Alice owns, the outgoing data can be audited to verify that no unexpected data is

escaping the boundaries of her PDS. In this way, rich applications and services can be built on top of the PDS that leverage all of these disparate data sources, while Alice still owns the underlying data behind these computations, and can take steps to preserve aspects of her privacy.

3.3.5 Key Research Questions

This vision is a world in which personal data that is easily available but yet the individual is protected. There are many technical challenges to accomplish this vision. For instance, the question-and-answer mechanism that allows certified answers to be shared instead of raw data requires the development of new privacy preserving technologies for user-centric on-the-fly anonymization.

Similarly, auditing the distribution and sharing of information in order to confirm that all data sharing is as intended requires the development of new algorithms and techniques to detect breaches and attacks.

There are also significant user interface questions, so that users really understand the risks and rewards they will be asked to opt into and are not overwhelmed with choices. A key idea for these interface questions is to use experimentation to determine user preferences for risk/reward, assessed via mechanisms such as differential privacy, in this question-answering environment.

3.3.6 Conclusion

As technologists and scientists, we are convinced that there is amazing potential in personal data, but also that the user has to be in control, making the trade-off between risks and benefits of data uses. openPDS is one attempt to provide a privacy-preserving Personal Data Store that makes it easy and safe for the user to own, manage and control his data. By anonymously just answering questions on-the-fly, openPDS opens up a new way for individuals to regain control over their data and privacy while supporting the creation of smart, data-driven applications.

3.4 openPDS/SafeAnswers[12]

3.4.1 Introduction

Personal metadata – digital information about users' location, phone call logs, or web-searches – is undoubtedly the oil of modern data-intensive science [129] and of the online economy [177]. This high-dimensional metadata is what allow apps to provide smart services and personalized experiences. From Google's search to Netflix's "movies you should really watch," from Pandora to Amazon, metadata is used by commercial algorithms to help users become more connected, productive, and entertained. In science, this high-dimensional metadata is already used to quantify the impact of human mobility on malaria [209] or to study the link between social isolation and economic development [94].

Metadata has however yet to realize its full potential. This data is currently collected and stored by hundreds of different services and companies. Such fragmentation makes the metadata inaccessible to innovative services, researchers, and often even to the individual who generated it in the first place. On the one hand, the lack of access and control of individuals over their metadata is fueling growing concerns. This makes it very hard, if not impossible, for an individual to understand and manage the associated risks. On the other hand, privacy and legal concerns are preventing metadata from being reconciled and made broadly accessible, mainly because of concerns over the risk of re-identification [79, 155, 194].

Here we introduce openPDS, a field-tested personal data store (PDS) allowing users to collect, store, and give fine-grained access to their metadata to third parties. We also introduce SafeAnswers, a new and practical way of protecting the privacy of metadata through a question and answer system. Moving forward, advancements in using and mining these metadata have to evolve in parallel with considerations of control and privacy [85, 163, 32, 175]. openPDS and SafeAnswers allow personal metadata to be safely shared and reconciled under the control of the individual.

[12]Published as de Montjoye Y.-A., Shmueli E., Wang S., Pentland A., openPDS: Protecting the Privacy of Metadata through SafeAnswers. PLoS One, 10.1371 (2014).

Towards Personal Data Stores

While questions of data ownership and the creation of repositories of personal data have been discussed for a long time [50, 204, 42, 153, 63, 119, 14, 20, 4], their deployment on a large-scale is a chicken-and-egg problem; users are waiting for compatible services while services are waiting for user adoption. Revelations of the collection and use of metadata by governments and companies [102, 110] have however recently drawn attention to their potential. The combination of 1) a public interest in questions of control but also use of their data, 2) political and legal support on data ownership [11, 22, 25, 17] and 3) the scale at which metadata can now be collected and processed, might trigger the large-scale deployment of PDS.

openPDS fully aligns with these trends. It uses the World Economic Forum definition of "ownership" of metadata [25]: the rights of possession, use, and disposal. It follows policies of the National Strategy for Trust Identities in Cyberspace (NSTIC) [22] and strongly aligns with the European Commission's reform of the data protection rules [11]. Finally, it recognizes that users are interacting with numerous data sources on a daily basis. Interoperability is thus not enough to achieve data ownership or address privacy concerns. Instead, openPDS implements a secure space acting as a centralized location where the user's metadata can live. openPDS can be installed on any server under the control of the individual (personal server, virtual machine, etc) or can be provided as a service (SaaS by independent software vendors or application service providers). This allows users to view and reason about their metadata and to manage fine-grained data access.

From an economic standpoint, data ownership by the individual fundamentally changes the current eco-system. It enables a fair and efficient market for metadata [177, 179] – a market where users can get the best services and algorithms for their metadata. Users can decide whether a service provides enough value for the amount of data it requests, and services can be rated and evaluated. Users are empowered to ask questions like "Is finding out the name of this song worth enough to me to give away my location?" Users can seamlessly give new services access to their

past and present metadata while retaining ownership. From a business standpoint, such data ownership is likely to help foster alternatives to the current data-selling and advertising-based business model. New business models focusing on providing hardware for data collection, storage for metadata, or algorithms for better using metadata might emerge while software for data collection and data management might be mostly open-source. The proposed framework removes barriers to entry for new businesses, allowing the most innovative algorithmic companies to provide better data-powered services [177].

Other approaches have been proposed for the storage, access control, and privacy of data. Previous approaches fall into two categories: cloud storage systems and personal data repositories. First, cloud storage systems, such as the ones that have been commercially developed by companies like Dropbox [10] and Carbonite [7], are a first approximation of a user-controlled information repository for personal data. They however focus on storing files and only implement the most basic type of access control, usually on a file or folder basis. They do not suggest any data aggregation mechanisms and, once access has been granted, the raw data is exposed to the outer world, potentially compromising privacy. Second, personal data repositories have been developed in academic [50, 204, 42, 153, 63, 119, 125, 39] and commercial settings [14, 20, 4]. All of these repositories are however restricted to specific queries on a particular type of data, such as interests or social security numbers. They provide only a basic access-control level, which means that once access to the data is authorized, privacy may be compromised. openPDS differs from previous approaches in its alignment with current political and legal thinking, its focus on large-scale metadata, and its SafeAnswers privacy-preserving mechanism.

On Privacy

There is little doubt that web searches, GPS locations, and phone call logs contain sensitive private information about an individual. In 2012, 72 percent of Europeans were already concerned about the use of their personal data [11]. The recent revelations are unlikely to have helped [110, 102]. Addressing users' legitimate privacy

concerns will soon be a prerequisite to any metadata usage.

Protecting the privacy of metadata is known to be a hard problem. The risks associated with high-dimensional metadata are often subtle and hard to predict and anonymizing them is known to be very hard. Over the last years, numerous works have exposed the risks of re-identification or de-anonymization of apparently anonymous datasets of metadata. An anonymous medical database was combined with a voters' list to extract the health record of the governor of Massachusetts [194] while the Kinsey Institute database was showed to be re-identifiable using demographics [184]. Twenty million web queries from around 650,000 AOL users were found to be potentially re-identifiable thanks to people's vanity searches [61] while the Netflix challenge dataset was de-anonymized using users' ratings on IMDB (The Internet Movie Database) [155]. Finally, mobility datasets of millions of users were found to be potentially re-identifiable using only four approximate spatio-temporal points [79].

Geospatial metadata, the second most recorded information by smartphone applications [199, 1], is probably the best example of the risks and rewards associated with metadata [190]. On the one hand, a recent report of the Electronic Frontier Foundation [59] worries about potentially sensitive information that can be derived from geospatial metadata. For example, geo-spatial metadata behavior collected from mobile phones has been shown to be very useful in predicting users' personalities [82]. On the other hand, the number of users of location-aware services, such as Yelp or Foursquare, are rising quickly as these services demonstrate their benefits to users.

Numerous ways of anonymizing personal data beyond the simple removal of explicit identifiers have been proposed. Similar to the original k-anonymity model [194], they aim minimize privacy risks while keeping data utility as high as possible. All anonymization models have however several major limitations.

Generic anonymization models have been designed for relatively low-resolution data and cannot be easily extended to high-dimensional data such as GPS location or accelerometer readings. Through generalization and suppression, k-anonymity makes every record in a given table indistinguishable from at least $k - 1$ other records, thereby making it impossible to identify an individual in that table. Variations and

alternatives include ℓ-diversity [140], which address attacks based on lack of diversity in the sensitive data and t-closeness [64, 133] which aims at maintaining the distribution of the sensitive data. The reader is referred to the surveys [35, 100] for further details. In metadata, any information that is unique to an individual can be used to re-identify him. Unicity ($\mathcal{E}$) has been used to quantify the re-identifiability of a dataset [79]. Most rich metadata datasets are expected to have a high $\mathcal{E}$. This means that, even if they are computationally tractable, generic privacy models are likely to result in most data having to be suppressed or generalized to the top-most values in order to satisfy the privacy requirement [34]. This curse of dimensionality led to the development of models dedicated to the anonymization of mobility data.

Mobility-focused anonymization models protect individual's mobility traces but only for very specific data applications or against specific re-identification attacks. The anonymization models in [53, 101, 216, 144, 173] protect the current location of the user, allowing him to anonymously perform accurate location-based searches. They however prevent any uses of historical metadata or side information, making them impractical for research and smart services using historical data. Other models [151, 198] allow for the anonymization of short successions of geospatial locations with no associated timestamps or [214] protect an individual's mobility data against re-identification at certain given times. These models however focus on anonymizing mobility data with a certain purpose or specific type of data in mind (i.e., current location, trajectory without timestamps or mobility data in given times). This makes these models impracticable for most data-science applications in academia and organizations.

Finally, all anonymization models, generic or mobility-focused, assume a setting in which the whole database is anonymized and published once. This makes it impractical, as (1) the same database is likely to be used to address different research questions (which might need specific pieces of information) and (2) smartphone applications or researchers might need access to the very latest pieces of information. Modifying existing anonymization models to support multiple releases has been shown to be non-trivial [181]. Indeed, anonymizing each publication on its own is not sufficient,

since a violation of privacy may emerge as a result of joining information from different publications. Anonymizing the whole database once and successively releasing the relevant part of the anonymized data is not a solution either, since newer data may become available. Several dedicated models were recently suggested to address the multiple publications setting [203, 62, 212, 181]. While very interesting, these models are based on extensions of the original one publication models and are thus very limited in the number and type of publications that they can handle.

SafeAnswers, a new paradigm

The goal of SafeAnswers is to turn an algorithmically hard anonymization and application-specific problem into a more tractable security one by answering questions instead of releasing copies of anonymized metadata.

Under the openPDS/SafeAnswers mechanism, a piece of code would be installed inside the user's PDS. The installed code would use the sensitive raw metadata (such as raw accelerometers readings or GPS coordinates) to compute the relevant piece of information within the safe environment of the PDS. In practice, researchers and applications submit code (the question) to be run against the metadata, and only the result (the answer) is sent back to them. openPDS/SafeAnswers is similar to differential privacy [91, 150], both being online privacy-preserving systems. Differential Privacy is however designed for a centralized setting where a database contains metadata about numerous individuals and answers are aggregate across these individuals. SafeAnswers is unique, as it focuses on protecting the privacy of a single individual whose data are stored in one place by reducing the dimensionality of the metadata before it leaves the safe environment. This individual-centric setting makes it practical for mobile applications or data-science researchers. It however introduces new privacy challenges [see Analysis].

Combined with openPDS, this simple idea allows individuals to fully use their data without having to share the raw data. SafeAnswers also allows users to safely grant and revoke data access, to share data anonymously without needing a trusted third-party, and to monitor and audit data uses [Fig. 3-1 and 3-2].

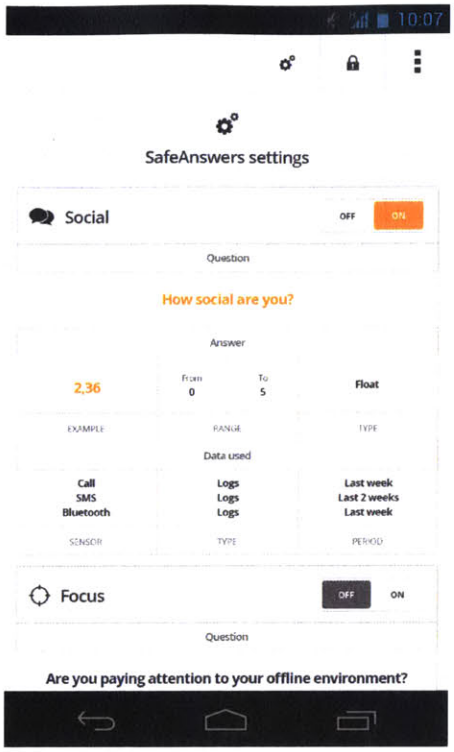


Figure 3-1: **Mockups of the proposed SafeAnswers settings presented to the user for approval. This screen shows the question answered, examples of the possible responses, and the sensors used to compute the response**

3.4.2 Results

The openPDS framework

The Dataflow

Looking at Fig. 3-3, consider a usecase in which a user uses a personalized music service such as PersonalizedMusic. Every time PersonalizedMusic needs to decide which song to play next on the user's mobile phone or desktop, it sends a request to the user's PDS. The actual computation of what song to play next is done by the PersonalizedMusic SafeAnswers module (SA module) inside the PDS front-end. As part of this processing, the PersonalizedMusic SA module accesses the back-end database in order to retrieve the required metadata. The PersonalizedMusic SA module would only access the raw metadata that it was authorized to when it was installed and all the processing would take place in a software sandbox. Upon completing its process-

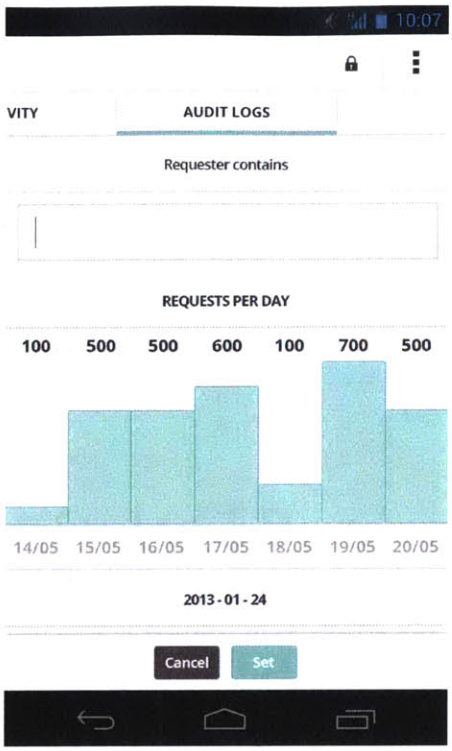


Figure 3-2: **Mockups of the proposed interface showing the number of requests sent by a given app per day.**

ing, the PersonalizedMusic SA module would return the name of the next song to play back to the front-end who will validate it and send it back to PersonalizedMusic.

The Database

Metadata are currently stored in a CouchDB database. CouchDB is a NoSQL store that stores data as a key to document mapping, where documents are JSON objects. CouchDB also provides a large range of existing functionality that lends itself well to the type of analysis needed to compute answers or reduce the dimensionality of the metadata. It has built-in support for MapReduce through CouchDB-Views, as well as data validation. All SafeAnswers modules share one unified database, and each SA module has a corresponding key prefix.

The Front-End

The front-end ensures that no unauthorized operations are carried out on the underlying metadata. SA modules are restricted to reading from the data sources

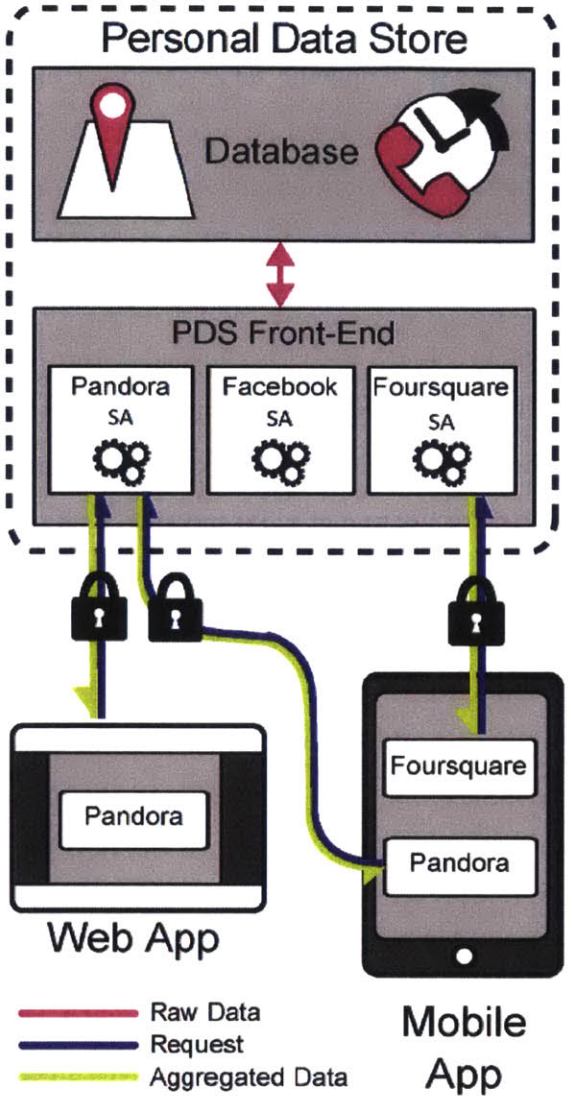


Figure 3-3: **openPDS system's architecture. LBSinc web or mobile app sent a request to the user's openPDS. The request is passed on to the LBSinc SA module, which requests access to the database in order to retrieve the metadata needed to compute the answer. The SA module computes the answer, which is then validated by the PDS Front-End and send back to the web or the mobile app**

they have explicitly listed as dependencies. CouchDB can also enforce access based on metadata types, time of access, time of collection, etc. The access control mechanism is implemented based on Django users and a permissioning system, where each app is registered as a user. We are working to decouple the access control mechanism and

the PDS using oAuth1.0 protocol [116]. This will allow an authentication server to hand out tokens associated with a specific service and set of metadata. In addition, SA modules are executed in a sandboxed environment, and all communications are encrypted using 256 bits SSL connections. In some implementations, PDSs can be managed from a web interface.

SafeAnswers is one key innovation of the openPDS framework. SafeAnswers allows for computations on user metadata to be performed within the safe environment of the PDS. Only safe answers, the exact information needed to provide the service, leave the PDS. SA modules are intimately tied to the notion of Design Documents in CouchDB. A CouchDB design document is intended to be a document that describes an application to be built on top of an underlying CouchDB instance. Each access of the SA module to the database has to be authorized and each SA module executes inside a sandbox. We are now working to add additional fields to the CouchDB design document specification to allow additional functionality, like SA module dependencies and permissions. These descriptions will be written in the SA module manifest to be programmatically enforced and to be presented to the user before installation.

In large-scale deployments, we expect that, instead of developing a SA module from scratch for each app, there will be common libraries that can be leveraged by SA modules or directly through a standard API. For example, there could be a library that supports functionality, like returning the current city a user is in [153], his radius of gyration in the past 7 days [108] or whether he is currently running. In the future, we also hope to further develop the SafeAnswers system to support sessions. This would allow for some of the most advanced data-science uses.

Field-studies and user feedback

Our two initial deployments offer a first qualitative evaluation of the system. The first field study is monitoring the daily behavior of individuals with diagnosed mental problems (PTSD, depression) and controls subjects for a month through their smartphones [168]. Data is used to reproduce the diagnoses of mental health conditions, focusing on changes in speech and social behavior. Recorded activities include psycho-motor activity, occupational activity, social interaction, and sleep behavior.

Fig. 3-4 presents "focus-group" results about the reaction of individuals to the openPDS framework ($N = 21$, 6 females and 15 males, median age category is 29 to 34 old). We consider the deployment to be a success, as 81% of individuals say they would use it in their personal life and, on a 1 to 5 scale (1: "Not at all comfortable" and 5: "Extremely comfortable"), are comfortable with the data collection (mean: 4, sem: 0.27). From a privacy perspective, we can see that the ability to delete data matters to participants (mean: 4.10, sem: 0.27). We can qualitatively see that users are generally comfortable sharing individual data with their primary care provider and mental health specialist. However, they seem to be less comfortable sharing such data with friends and potentially their family members. We can also see that anonymity matters to participants (mean:4 sem:0.30) and that they are significantly more comfortable sharing anonymous, rather than individual, data (p-value < 0.005 with a one-tailed, paired, non-parametric Kolmogorov-Smirnov test on 4 specific sharing questions, and mean:4 sem:0.25 when asked on the importance of anonymizing shared data). All these emphasize the relevance of the openPDS/SafeAnswers framework.

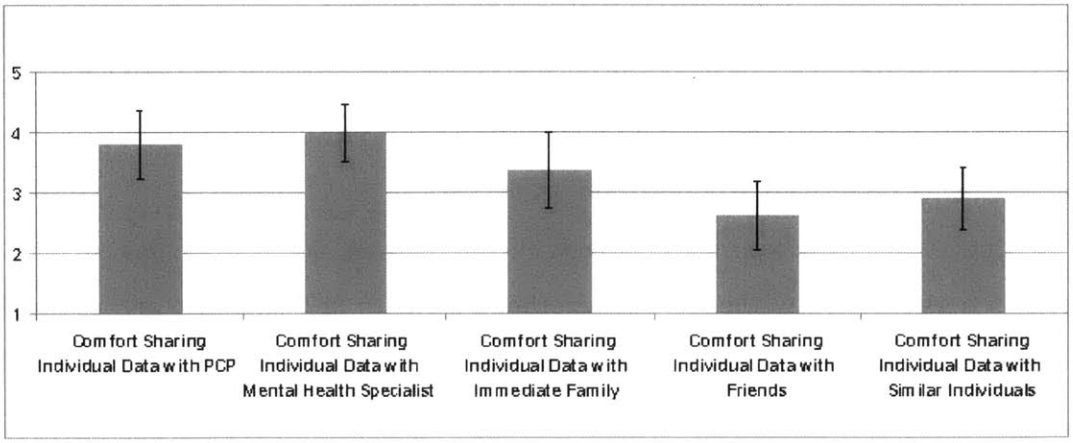


Figure 3-4: **Individuals' reaction to data sharing. The error bars are bootstrapped 95% high-density intervals. We can qualitatively see that users are generally comfortable sharing individual data with their primary care provider and mental health specialist. They however seem to be less comfortable sharing such data with friends and potentially family members.**

A second study, the mobile territorial lab, in partnership with Telecom Italia, Telefonica, and the Fondazione Bruno Kessler, is now underway. It is composed of 70

young parents living in Trento and its premises. The aim here is to create a long-term living lab to study user behavior and to perform user studies. Participants' behavior is recorded using an extended version of the open-sensing framework FunF [13]. All collected metadata are stored on users' PDSs.

3.4.3 Discussion

Performance

openPDS may introduce a performance overhead caused by its distributed nature, the added security and privacy mechanisms and the group computation mechanism [see Analysis].

First, the distributed nature of openPDS requires services to access the user's PDS when an answer has to be computed. In cases where computing the answer is fast, the latency it imposes might make an openPDS-based solution impracticable. Solutions such as precomputing some values and locally caching them might help. However, in cases where computing the answer inside the PDS dominates the total execution time, this might not significantly impact the user experience. In fact, this might actually introduce a performance boost, since it parallelizes the computations that are being performed at a per-PDS level.

Second, the added security and privacy mechanisms described below may also result in performance overhead. This overhead needs to be taken into account when choosing the appropriate mechanism. For example, the on-the-fly nature of openPDS/SafeAnswers may lead to inference of sensitive data if the results of several queries are joined together. On the one hand, using techniques such as the one suggested by [181] may be very efficient in preventing such inference, but they are relatively expensive in computation time. On the other hand, adding noise to query results may not be equally efficient, but would result in a much faster computation time. Advanced techniques might thus be crucial when dealing with credit card or location data, but noise addition might be sufficient to protect less sensitive data such as accelerometer readings.

For many years, group computation has been of theoretical interest only. Great improvements and actual field-studies in domains such as electronic voting, actioning, and data mining have recently made group computation–also called Secure Multiparty Computation, or SMC–of practical interest [161]. Similar to network latency, the overhead of SMC might become reasonable if computing the answer dominates the total computation time. SMC has furthermore recently been generalized into belief propagation algorithms [126]. This means that every node of the computation does not have to communicate with every other anymore, thereby reducing the overhead.

Usage Experience

In this section we describe two short scenarios for a user and a developer switching to an openPDS/SafeAnswers system for mobile applications.

End-User Suppose Alice wants to install and use a smartphone app like LBSinc, a location-based check-in application, without using a PDS. Alice downloads the app onto her phone, authorizes LBSinc to access her phone's network communication and GPS coordinates, and creates a user account with LBSinc. The LBSinc app starts collecting metadata about her and stores it all in its back-end servers. Under this model it is difficult for Alice to access the metadata LBSinc uses to makes inferences about her, or to remove the metadata she does not want LBSinc to access or use.

Alternatively, Alice could decide to download a PDS-aware version of LBSinc. She installs it just like she would install any other smartphone app and authorizes it to access only her phone's network communication. When used for the first time, the smartphone app prompts her to enter her PDS URI. Alice then sees exactly what metadata the LBSinc SA module will have access to and examples of the answers [see Fig. 2], the relevant summarized information that will be send back to LBSinc. If she accepts, the LBSinc SA module is installed onto her PDS and she can start using it.

App Developer Suppose a developer now wants to implement MyMusic, a smartphone app that plays music to Alice based on her preferences and current activity. Under the current model, he would first have to develop a smartphone app to col-

lect the metadata on Alice's phone, record it, and periodically send it to a server. He would then develop a server with an internal database to store the raw activity data he collects, a secured API for this database to receive the metadata, and a way to anonymize the metadata or at least separate the user account information from the metadata. He could then start developing an algorithm to decide which song or type of music to play. The initial picture he would have of users would be very rough, as he would have no prior metadata to work with. Finally, he would have to wait to collect a sufficient amount of metadata before being able to provide adequate recommendations.

If operating within the openPDS/SafeAnswers framework, the metadata that the developer needs are likely to have already been collected either by a metadata collection app [5] or by another application or service. The developer would then spend most of his time writing an SA module that would decide which song or type of music to play and test it on development copies of PDSs. The PDS front-end would take care of creating the API and of securing the connection for him. The developer's algorithm would be able to access a potentially large set of metadata, including historical metadata.

3.4.4 Analysis

The openPDS framework suggests several mechanisms for enhancing the privacy and security of personal metadata: SafeAnswers, access control, sandboxes, and network encryption. In this section, we discuss several cases where these might fall short and discuss potential counter-measures.

Protecting aggregated answers of groups

A practical example would be a service, such as CouponInc, which wants to execute a simple query to know how many of its users are around a certain shop to send them a special coupon. CouponInc might want to issue a query like "How many users are in this geographical area at the current time?" or "How active are these users during

lunch time?"

In a distributed setting, such computation falls under the well-studied field of secure multi-party computation (SMC) [105], where the querying agent never sees any individual user's metadata but can access information aggregated across users. User privacy is preserved, as each PDS only sends cryptographically masked messages to other nodes in the network.

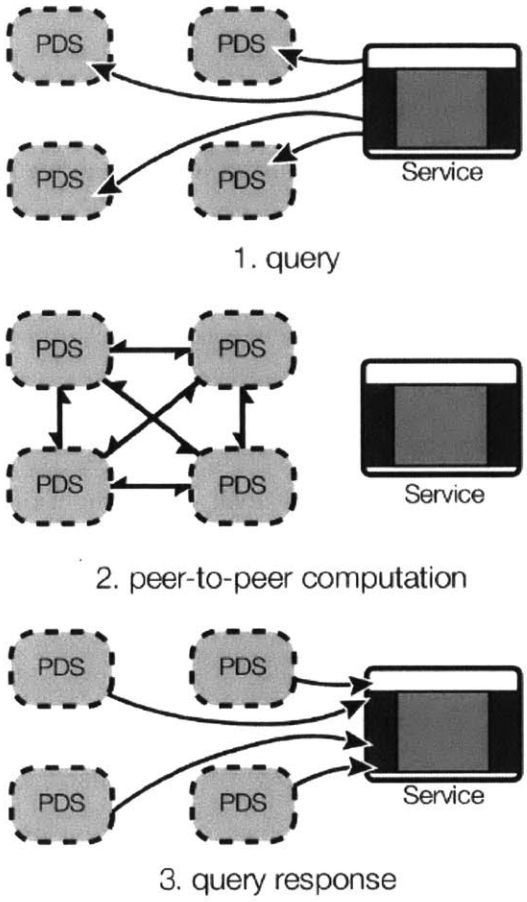


Figure 3-5: **Group Computation Overview.** (1) A querying agent (like Coupon-Inc) passes a function that its wants a collaborative answer for, along with a list of URI to PDSs. (2) PDSs all trade messages in order to compute a collaborative answer. (3) The answer is reported back to the querying agent

Such a cryptographic technique fits elegantly into the PDS model of computation [Fig. 3-5]. Rather than anonymizing and computing over-complex data items, like

GPS coordinates, the SA modules could compute features locally to each user's PDS, reducing the dimensionality of the metadata. After the local computation is done, the inferred facts–e.g. whether or not a given user is in a given geographical area–can be aggregated in a privacy-preserving way. This means that even the low-dimension answer cannot be associated with a particular user.

Attacks in the case of well-behaved apps

Even in the absence of attackers, apps that behave as they are supposed to might pose a risk to users' privacy. We notice two major challenges: (1) How can an openPDS/SafeAnswers determine the required level of aggregation given that it only has access to the metadata of a single user? (2) Well-behaved apps could inadvertently collect data whose combinations may allow others to infer sensitive information.

A potential solution to the first challenge might be found in [79]. The authors studied fifteen months of human mobility data for one and a half million individuals, and found that one formula determines the uniqueness of an individual's mobility traces, given the traces' resolution (i.e., level of aggregation) and the amount of background knowledge available to the adversary. If extended to other types of data, such an equation could be used by SafeAnswers to determine the required level of aggregation needed when answering a query.

The fields of Privacy Preserving Data Publishing and Mining aim to address a problem similar to the second challenge: how to anonymize the current publication of a database so that the combination of all past and current anonymized publications respect privacy. These works suggest several interesting assumptions and techniques that could be adopted by the openPDS/SafeAnswers framework. For example, the authors of [181] show that the problem of accurately calculating the level of privacy imposed by a set of three or more publications is NP-hard. The authors then suggest a relaxed method for calculating the privacy level in polynomial time. Their method is based on joining the set of publications into a single table, which can then be checked against some privacy requirement. They also suggest a supplementing algorithm for anonymizing the current publication so that the required privacy level is obtained.

Their methods might be used by SafeAnswers in order to determine whether the current set of queries and potential future queries might compromise privacy.

Work in statistical databases might also help address the second challenge [210]. A statistical database aims to allow the execution of statistical queries without compromising the confidentiality of any individual represented in the database. Two approaches used in this field could be useful for SafeAnswers: (1) A query restriction rejects each query that could compromise a user's privacy and provides accurate answers to legitimate queries. The computation of what is a legitimate query is usually based on the size of the query's results or the extent of overlap between queries. Note however that the denial of a query may, in itself, provide information to an an attacker. (2) Perturbation gives approximate answers by adding noise to the answers computed from the original metadata. Regardless of the specific perturbation technique, the designer must attempt to produce statistics that accurately reflect the underlying database. Such perturbed answers might however not be acceptable for all uses.

Attacks in the case of malicious apps

While well-behaved apps might inadvertently collect sensitive information, apps that are voluntarily not playing by the rules pose a serious threat to user's privacy. The major risk we see here is how to protect the metadata against an app that deliberately tries to infer sensitive information by over-querying a user's openPDS or by colluding with other apps.

Technically, numerous techniques from anomaly detection may help SafeAnswers detect suspicious behavior. For example, a service that suddenly changes its query pattern; querying for location every minute while it used to ask user's location and speed a few times in a row 3 times a day. The detection of anomalies, outliers, or rare events, has recently gained a lot of attention in many security domains, ranging from video surveillance and security systems to intrusion detection and fraudulent transactions. Accordingly [66], most anomaly detection methods are based on the following techniques: classification, nearest neighbor, clustering, statistical, information theo-

retic, and spectral. Any of these techniques, or their combination, can potentially be used by SafeAnswers.

Anomaly detection could also be combined with reputation systems to allow for a group of openPDSs to exchange information about modules and services in real-time. The P2P reputation systems literature considers different types of malicious behavior that can be blocked with the help of reputation systems. These give us a foretaste of potential risks. "Traitors" are services who initially behave properly but then start to misbehave and inflict damage on the community. This technique is particularly effective when the service has become respectable and well installed. "Whitewashers" are services who leave and rejoin the system with new identities in order to purge the bad reputations they acquired under their previous identities. Finally, "Collusions" are a group of malicious services acting together to cause damage. Such reputation systems could be combined with other privacy mechanisms discussed here. For example, an openPDS might decide to allow a service with a medium rating to execute restricted or noisy queries but temporarily block a service whose rating suddenly dropped.

Various UI mechanisms can also be used to warn users of potentially malicious apps before they are installed. For example, trust could be used to rate service providers. Adapting the definition from [152], trust would reflects a user's or a PDS's subjective view of a service, while reputation could be considered a collective measure of trust reflecting a group view of that service. Work by [130] shows that the reputation of the service provider matters more than the specific data being accessed and hints at the potential usefulness of a reputation system to help users decide which services to trust. Various principles for computing reputation and trust can be found in [124]. Besides a simple summation or average of ratings, the authors mention discrete models in which trust is a discrete value from a predefined set of values, fuzzy models, bayesian systems, belief models, and flow models.

Attacks compromising the host

Finally, openPDS is vulnerable to the traditional security and privacy issues of any hosted system. Attackers could compromise the authentication/control mechanisms

or impersonate existing users to gain access to the database or to corrupt the system. For instance, in the case of virtual machines hosting openPDSs, an attacker's virtual machine can legitimately be located in the same physical machine as openPDSs virtual machines. This is, however, not specific to openPDS, and similar issues exist with any hosted systems, such as SaaS, virtual machine and traditional servers. Solutions include hypervisors [169] or data-at-rest encryption [182, 170] such as homomorphic encryption schemes [103]. The main difference openPDS introduces is having the data distributed across machines, systems, and implementations of openPDS. While a full analysis is beyond the scope of this paper, one might imagine that a distributed and heterogeneous system might be harder to attack than some of the traditional centralized ones especially if information is shared across machines [see previous section].

3.4.5 Conclusion

Finally, as technologists and scientists, we are convinced that there is an amazing potential in personal metadata, but also that benefits should be balanced with risks. By reducing the dimensionality of the metadata on-the-fly and actively protecting users, openPDS/SafeAnswers opens up a new way for individuals to regain control over their privacy.

openPDS/SafeAnswers however still face a number of challenges. Each challenges includes several potential directions for future research: (1) the automatic or semi-automatic validation of the processing done by a PDS module; (2) the development of SafeAnswers privacy-preserving techniques at an individual level for high-dimensional and ever-evolving data (mobility data, accelerometer readings, etc.) based on existing anomaly detection framework and potentially stored in highly-decentralized systems; (3) the development or adaptation of privacy preserving data-mining algorithms to an ecosystem consisting of distributed PDSs; and (4) UIs allowing the user to better understand the risks associated with large-scale metadata and to monitor and visualize the metadata used by applications.

Conclusion

The results of this thesis assert the need to, once again [206], deeply rethink our approach to data protection in order to keep up with the evolution of technology. Our ability to collect and process large amounts of data has greatly increased in the last decade. An equivalent paradigm-shift in our ability to protect data is now required to provide the level of privacy needed for the harmonious development of our societies.

This requires us to acknowledge the limits of the traditional de-identification model[13] and to refocus our policies on the original notion and intention of anonymity. Technically, we show that privacy-through-security approaches have great potential to help ensure anonymity and protect the privacy of individuals in the age of big data.

This thesis first argues that the premise of data anonymization, that someone can "hide in the crowd", is inadequate to protect the privacy of individuals in modern high-dimensional datasets (data from mobile phones, the Internet of Things, public transportation, wearables, etc). We introduced *unicity* and used it to show that only a few points are needed to uniquely identify an individual with high likelihood in both mobile phone [79] and credit card datasets [83]. We furthermore showed that anonymization strategies, such as data generalization, are not sufficient to ensure anonymity in high-dimensional datasets. Our work has been replicated on four different mobile phone datasets: Two Italian [65] and one Latin American [172] datasets,

[13]We here use "de-identification model" to refer to the release of datasets that have been anonymized (or de-identified) and undergone risk assessment

and one dataset from an unnamed country [187]. All reached the same conclusion: that mobile phone data are high-unicity.

This thesis goes on to show that the second pillar of de-identification, risk assessment, is similarly crumbling. As datasets become richer, adequate risk assessment will need to consider not only what is directly visible about an individual in the data, but also what an algorithm could uncover from the data, now or in the future. For example, using mobile phone data we showed that machine learning algorithms could predict the personality traits of an individual up to 1.7 times better than random [82]. In the age of big data, assessing the risk of inference requires significant investments in specialized training datasets and fast-evolving machine learning techniques [87]. Comprehensive risks assessments will become increasingly difficult to perform, ultimately strongly limiting their relevance.

Taken together, the limits of anonymization [79, 83, 155, 45, 115, 196] and risk of inference [82] in high-dimensional datasets strongly restrict the pertinence of the de-identification model. The de-identification model, where control over the data is effectively lost [14] is no longer a useful basis for policy[15]

- As we have shown, the removal or absence of legally defined "Personally Identifiable Information (PII)" is not an effective anonymization method. It should not be considered enough to make data "non-personal" [131] and release it free of legal protections (the de-identification model).

- More advanced methods such as data generalization, sampling, suppression, or noise addition can sometimes help limit the risk of re-identification [80]. These methods are however still insufficient to classify data as "non-personal" and release it free of legal protections (the de-identification model)[16]. Even worse,

[14]Data Use and Non-Disclosure Agreements are notoriously difficult to enforce across jurisdictions and are only practical when data is shared with a handful of trusted partners. Released data such as the AOL search dataset and the NYC taxi cab dataset are still available online despite known, well-documented vulnerabilities [37, 165]

[15]A view since shared by the President's Council of Advisors on Science and Technology who concluded that "Anonymization remains somewhat useful as an added safeguard, but it is not robust against near-term future re-identification methods. PCAST does not see it as being a useful basis for policy." [159]

[16]Note that here we focus on individual-level data that have been anonymized using the advanced

such refined methods might be counter-productive by giving a false sense of safety that it would be very hard to re-identify the data.

- Similarly, estimated risks of re-identification are also not a useful basis for policy. Indeed, as we argued before "one can always [...] artificially lower the estimated likelihood of reidentification through the use of arbitrary and debatable assumptions" [81], even without any changes made to the actual dataset [46]. For instance, pre-release estimations of the risk of re-identification for the Heritage Health Prize dataset ranged from .0084% under liberal assumptions [96] to 12.5% under more conservatives ones [154].

The inadequacy of the de-identification model been resisted by some who have argued that it would result in a "tragedy of the data commons" [214] or that this is some kind of "inevitable trade-off" [47]. Equating the recognition of the shortcomings of the de-identification model with a decrease in data use, e.g. for research, is misleading. Firstly, it assumes that the de-identification model is the only one available, effectively disregarding existing and applicable alternatives including modern privacy-through-security approaches. Secondly, it ignores the fact that the scope of the privacy laws in the United States and Europe is much broader than just the use of data for research including, e.g. the selling of individual-level anonymized data.

Recognizing the limits of the de-identification model, three main solutions exist from a policy perspective. First, we could consider all individual-level data (not aggregated) to be personal. Their use and collection would be governed by existing privacy laws. This, however, would give no incentive to take the important, but insufficient step of removing names and other direct identifiers when collecting or using data, often unnecessarily increasing risks to individual privacy.

The second option would be to create a third category of data: *identifiable* data [179] for which "some non-remote possibility of future identification" exists. Data could be *identified* if it contains direct identifiers (e.g. names - currently personal data), *non-identifiable* (currently not personal data), or *identifiable.* In the

methods mentioned above. This is not necessarily applicable to new–often much smaller scale–extracted data e.g. behavioral indicators [77] or "answers"

latter case, the authors propose that some, but not all, legal protections would apply. Aggregated data can probably usually be considered non-identifiable data and the potential of big data is in individual-level (often high dimensional) data. As with the current de-identification approach, this assumes that a useful and defensible line can be technically drawn between non-identifiable and identifiable individual-level data. As we have discussed above, estimated risks of re-identification are not a useful basis for policy.

The best option is probably to keep, from a policy standpoint, the conceptual notion of anonymity while ensuring its promise through legal and privacy-through-security means. The data owner guarantees that data will always only be used in aggregate form (e.g. the results of statistical models), data will not be re-identified nor reconciled across datasets (data merge), data will not be used to make decisions about an individual, etc [17].

From a technical perspective, we argue that privacy-through-security model can strongly help data owners ensure the promise of anonymity. Here, as opposed to the de-identification model, raw data is never shared. Third parties are given a remote access to the data, and data access and use are controlled to help ensure that the data is used appropriately. Four main classes of privacy-through-security models exist, depending on the sensitivity of the data and the stage of development of the research question:

1. Remote data access [113]: direct identifiers are removed and data is protected through access control, IP-restrictions, and active data monitoring.

2. Question-and-answers systems [78]: only answers computed from the raw data [77] are shared with third parties.

3. Aggregated answers: answers are aggregated across individuals, e.g. using secure multi-party computation [105], before being shared.

[17] Such more conceptual approach is similar to the proposed European privacy directive and the US "use-based" strategy.

4. Differential privacy [91]: answers are aggregated across individuals and sufficient noise is added to give formal privacy guarantees.

We believe that anonymity provides a clear and easy to understand promise for data privacy and a useful basis for policy. Once disconnected from the de-identification model, anonymity can truly be ensured using legal and privacy-through-security strategies that can adapt to technological evolutions.

Bibliography

[1] The app genome project. http://blog.myLookout.com/. Accessed: 2011-07-27.

[2] Apple privacy policy. http://www.apple.com/legal/privacy/. Accessed: 2011-07-25.

[3] Apple's app store downloads top 25 billion. http://www.apple.com/pr/library/2012/03/05Apples-App-Store-Downloads-Top-25-Billion.html. Accessed: 2012-03-28.

[4] Azigo website. https://www.azigo.com/.

[5] Behav.io website. http://behav.io/.

[6] Boom! foursquare crosses 2 million users. http://techcrunch.com/2010/07/10/foursquare-crosses-2-million-users/. Accessed: 2010-08-25.

[7] Carbonite Backup website. http://www.carbonite.com/en/.

[8] D4D-Cote d'Ivoire - Book of abstract. http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf. [Online; accessed 16-July-2014].

[9] Draft african union convention on the establishment of a credible legal framework for cyber security in africa. http://www.au.int/en/cyberlegislation.

[10] Dropbox website. http://www.dropbox.com/.

[11] European Commission proposes a comprehensive reform of data protection rules to increase users' control of their data and to cut costs for businesses. http://europa.eu/rapid/pressReleasesAction.do?reference=IP/12/46&format=HTML&aged=0&language=EN&guiLanguage=en.

[12] 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: 2012-07-01.

[13] FunF website. http://www.funf.org/.

[14] Higgins website. http://www.eclipse.org/higgins/.

[15] In us cities, open data is not just nice to have; it's the norm. http://www.theguardian.com/local-government-network/2013/oct/21/open-data-us-san-francisco. Accessed: 2015-01-30.

[16] International strategy for cyberspace. http://www.whitehouse.gov/sites/default/files/rss_viewer/internationalstrategy_cyberspace.pdf. Accessed: 2015-01-30.

[17] International Strategy for Cyberspace. http://www.whitehouse.gov/sites/default/files/rss_viewer/internationalstrategy_cyberspace.pdf.

[18] MIT City Science. http://cities.media.mit.edu/. [Online; accessed 16-July-2014].

[19] 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: 2012-07-17.

[20] Mydex website. http://mydex.org/.

[21] National strategy for trust identities in cyberspace. http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf. Accessed: 2015-01-30.

[22] National Strategy for Trust Identities in Cyberspace. http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf.

[23] Personal data: The emergence of a new asset class. http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf. Accessed: 2015-01-30.

[24] Reality mining of mobile communications: Toward a new deal on data. https://members.weforum.org/pdf/gitr/2009/gitr09fullreport.pdf. Accessed: 2015-01-30.

[25] Reality Mining of Mobile Communications: Toward a New Deal on Data. https://members.weforum.org/pdf/gitr/2009/gitr09fullreport.pdf.

[26] Skyhook wireless spotrank overview. http://www.skyhookwireless.com/\location-intelligence/. Accessed: 2012-07-17.

[27] Us consumer privacy bill of rights. http://www.whitehouse.gov/sites/default/files/privacy-final.pdf. Accessed: 2015-01-30.

[28] 2013 federal reserve payments study. http://www.frbservices.org/files/communications/pdf/research/2013_payments_study_summary.pdf, 2013.

[29] The trust advantage: How to win with big data. http://www.bcgperspectives.com/content/articles/information_technology_strategy_consumer_products_trust_advantage_win_big_data/, 2013.

[30] Call for help. http://www.economist.com/news/leaders/21627623-\mobile-phone-records-are-invaluable-tool-combat-ebola-they-\should-be-made-available, 2014.

[31] Waiting on hold. http://www.economist.com/news/science-\and-technology/21627557-mobile-phone-records-would\-help-combat-ebola-epidemic-getting-look, 2014.

[32] H. Abelson, K. Ledeen, and H. Lewis. *Blown to bits: your life, liberty, and happiness after the digital explosion.* Addison-Wesley Professional, 2008.

[33] Alessandro Acquisti, Curtis R Taylor, and Liad Wagman. The economics of privacy. *Available at SSRN 2580411*, 2015.

[34] C.C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, pages 901–909. VLDB Endowment, 2005.

[35] C.C. Aggarwal and P.S. Yu. *Privacy-preserving data mining: models and algorithms.* Springer-Verlag New York Inc, 2008.

[36] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, 2011.

[37] Michael Arrington. Aol proudly releases massive amounts of private data. http://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/.

[38] Sonia M Arteaga, Mo Kudeki, and Adrienne Woodworth. Combating obesity trends in teenagers through persuasive mobile technology. *ACM SIGACCESS Accessibility and Computing*, (94):17–25, 2009.

[39] Mark Assad, David J Carmichael, Judy Kay, and Bob Kummerfeld. Personisad: Distributed, active, scrutable model framework for context-aware services. In *Pervasive Computing*, pages 55–72. Springer, 2007.

[40] Massachusetts Bay Transportation Authority. Real-time commuter rail data. http://www.www.mbta.com/rider_tools/developers/default.asp?id=21899. Accessed: 2015-01-30.

[41] Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, 2010.

[42] R. Baden, A. Bender, N. Spring, B. Bhattacharjee, and D. Starin. Persona: an online social network with user-defined privacy. In *ACM SIGCOMM Computer Communication Review*, volume 39, pages 135–146. ACM, 2009.

[43] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[44] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[45] Michael Barbaro, Tom Zeller, and Saul Hansell. A face is exposed for aol searcher no. 4417749. *New York Times*, 9(2008):8For, 2006.

[46] Daniel Barth-Jones, Khaled El Emam, Jane Bambauer, Ann Cavoukian, and Bradley Malin. Assessing data intrusion threats. *Science (New York, NY)*, 348(6231):194, 2015.

[47] Daniel C. Barth-Jones. Does de-identification work or not? `http://www.fiercebigdata.com/story/does-de-identification-work-or-not/2014-06-23`.

[48] D Bates and D Watts. Nonlinear regression analysis and its applications. *John Wilex & Sons*, 1988.

[49] Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 217–228. IEEE, 2005.

[50] G. Bell. A personal digital store. *Communications of the ACM*, 44(1):86–91, 2001.

[51] Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan Von Schreeb. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti. *PLoS medicine*, 8(8):1128, 2011.

[52] Michael J Benson and John P Campbell. To be, or not to be, linear: An expanded representation of personality and its relationship to leadership performance. *International Journal of Selection and Assessment*, 15(2):232–249, 2007.

[53] A.R. Beresford and F. Stajano. Location privacy in pervasive computing. *Pervasive Computing, IEEE*, 2(1):46–55, 2003.

[54] Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli, and Marco Luca Sbodio. Allaboard: a system for exploring urban mobility and optimizing public transport using cellphone data. In *Machine Learning and Knowledge Discovery in Databases*, pages 663–666. Springer, 2013.

[55] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J Christopher Westland. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602–613, 2011.

[56] Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *arXiv preprint arXiv:1210.0137*, 2012.

[57] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[58] Theodora Bloom, Emma Ganley, and Margaret Winker. Data access for the open access literature: Plos's data policy. *PLoS Biol*, 12(2):e1001797, 2014.

[59] Andrew J Blumberg and Peter Eckersley. On locational privacy, and how to avoid losing it forever. *Electronic frontier foundation*, 10(11), 2009.

[60] Geoffrey Boulton. Open your minds and share your results. *Nature*, 486(7404):441–441, 2012.

[61] D. Butler. Data sharing threatens privacy. *Nature*, 449(7163):644, 2007.

[62] Ji-Won Byun, Yonglak Sohn, Elisa Bertino, and Ninghui Li. Secure anonymization for incremental datasets. In *Secure Data Management*, pages 48–63, 2006.

[63] R. Cáceres, L. Cox, H. Lim, A. Shakimov, and A. Varshavsky. Virtual individual servers as privacy-preserving proxies for mobile devices. In *Proceedings of the 1st ACM workshop on Networking, systems, and applications for mobile handhelds*, pages 37–42. ACM, 2009.

[64] J. Cao, P. Karras, P. Kalnis, and K.L. Tan. Sabre: a sensitive attribute bucketization and redistribution framework for t-closeness. *The VLDB Journal*, 20(1):59–81, 2011.

[65] Alket Cecaj, Marco Mamei, and Franco Zambonelli. Re-identification and information fusion between anonymized cdr and social network data. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–14, 2015.

[66] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[67] Segolene Charaudeau, Khashayar Pakdaman, and Pierre-Yves Boëlle. Commuter mobility and the spread of infectious diseases: application to influenza in france. 2014.

[68] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 17(3):433–450, 2013.

[69] Michael T Clanchy. *From memory to written record: England 1066-1307*. Harvard University Press, 1979.

[70] Aaron Clauset and Nathan Eagle. Persistence and periodicity in a dynamic proximity network. *Proc. DIMACS*, 2007.

[71] CNET. 2011 ends with almost 6 billion mobile phone subscriptions. `http://news.cnet.com/8301-1023_3-57352095-93/2011-ends-with-almost-6-billion-mobile-phone-subscriptions/`. Accessed: 2015-01-30.

[72] CNN. Your phone company is selling your personal data. `http://money.cnn.com/2011/11/01/technology/verizon_att_sprint_tmobile_privacy/index.htm`. Accessed: 2015-01-30.

[73] European Commission. General data protection regulation. `http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf`, 2012.

[74] Scott Counts and Kristin Brooke Stecher. Self-presentation of personality during online profile creation. In *ICWSM*, 2009.

[75] Jeffrey M Cucina and Nicholas L Vasilopoulos. Nonlinear personality–performance relationships and the spurious moderating effects of traitedness. *Journal of Personality*, 73(1):227–260, 2004.

[76] Y-A de Montjoye. Privacy-tools. `https://github.com/yvesalexandre/privacy-tools/`.

[77] Y-A de Montjoye, Luc Rocher, and Sandy Pentland. bandicoot, a python toolbox to extract behavioral indicators from metadata. `http://bandicoot.mit.edu/`.

[78] YA de Montjoye, E Shmueli, SS Wang, and AS Pentland. openpds: Protecting the privacy of metadata through safeanswers. *PloS one*, 9(7):e98790, 2014.

[79] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Nature S.Rep.*, 3, 2013.

[80] Yves-Alexandre de Montjoye, Jake Kendall, and Cameron F Kerry. Enabling humanitarian use of mobile phone data. 2014.

[81] Yves-Alexandre de Montjoye and Alex Sandy Pentland. Assessing data intrusion threats–response. *Science (New York, NY)*, 348(6231):195–195, 2015.

[82] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Sandy Pentland. Predicting personality using novel mobile phone-based metrics. In *Proc. SBP*, pages 48–55. Springer, 2013.

[83] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, et al. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

[84] Yves-Alexandre de Montjoye, Zbigniew Smoreda, Romain Trinquart, Cezary Ziemlicki, and Vincent D Blondel. D4d-senegal: the second mobile phone data for development challenge. *arXiv preprint arXiv:1407.4885*, 2014.

[85] Yves-Alexandre de Montjoye, Samuel S Wang, Alex Pentland, Dinh Tien Tuan Anh, Anwitaman Datta, et al. On the trusted use of large-scale personal data. *IEEE Data Eng. Bull.*, 35(4):5–8, 2012.

[86] Rodrigo de Oliveira, Alexandros Karatzoglou, Pedro Concejero Cerezo, Ana Armenta Lopez de Vicuña, and Nuria Oliver. Towards a psychographic user model from mobile phone usage. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 2191–2196. ACM, 2011.

[87] Li Deng and Dong Yu. Deep learning: methods and applications. *Foundations and Trends in Signal Processing*, 7(3–4):197–387, 2014.

[88] Duane DeSieno. Adding a conscience to competitive learning. In *Neural Networks, 1988., IEEE International Conference on*, pages 117–124. IEEE, 1988.

[89] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, and Andrew J. Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 2014.

[90] Trinh-Minh-Tri Do and Daniel Gatica-Perez. By their apps you shall understand them: mining large-scale patterns of mobile phone usage. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, page 27. ACM, 2010.

[91] Cynthia Dwork. Differential privacy. In *Automata, languages and programming*, pages 1–12. Springer, 2006.

[92] E-ZPass. About us - statistics.

[93] Nathan Eagle, Y-A de Montjoye, and Luís Bettencourt. Community computing: Comparisons between rural and urban societies using mobile phone data. In *Computational Science and Engineering*, volume 4, pages 144–150. IEEE, 2009.

[94] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[95] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[96] Khaled El Emam, Luk Arbuckle, Gunes Koru, Benjamin Eze, Lisa Gaudette, Emilio Neri, Sean Rose, Jeremy Howard, and Jonathan Gluck. De-identification methods for open health data: the case of the heritage health prize claims dataset. *Journal of medical Internet research*, 14(1), 2012.

[97] eMarketer. Us mobile payments to top $ 1 billion in 2013. `http://www.emarketer.com/Article/US-Mobile-Payments-Top-1-Billion-2013/1010035`, 2013.

[98] Kendall et al. Using mobile data for development. `http://www.impatientoptimists.org/Posts/2014/07/Big-Data-and-How-it-Can-Serve-Development`, 2014.

[99] Office of Management & Budget Executive Office of The President. Safeguarding against and responding to the loss of personal information, memorandum m-07-16. (May 22, 2007).

[100] B. Fung, K. Wang, R. Chen, and P.S. Yu. Privacy-preserving data publishing: a survey of recent developments. *ACM Computing Surveys (CSUR)*, 42(4):1–53, 2010.

[101] B. Gedik and L. Liu. Location privacy in mobile systems: A personalized anonymization model. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on*, pages 620–629. Ieee, 2005.

[102] Barton Gellman and Ashkan Soltani. NSA tracking cellphone locations worldwide, snowden documents show. *The Washington Post*, 2013.

[103] C. Gentry. *A fully homomorphic encryption scheme.* PhD thesis, Stanford University, 2009.

[104] Jim Giles. Making the links. *Nature*, 488(7412):448–450, 2012.

[105] O. Goldreich. Secure multi-party computation. *Manuscript. Preliminary version*, 1998.

[106] Philippe Golle and Kurt Partridge. On the anonymity of home/work location pairs. In *Pervasive computing*, pages 390–397. Springer, 2009.

[107] André Gomez and Rapson Gomez. Personality traits of the behavioural approach and inhibition systems: Associations with processing of emotional stimuli. *Personality and Individual Differences*, 32(8):1299–1316, 2002.

[108] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[109] Peter Gould. Letting the data speak for themselves. *Annals of the Association of American Geographers*, 71(2):166–176, 1981.

[110] Glenn Greenwald and Ewen MacAskill. NSA prism program taps in to user data of apple, google and others. *The Guardian*, 2013.

[111] Stephen Grossberg. Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological cybernetics*, 23(3):121–134, 1976.

[112] Miniwatts Group. World internet users and population stats. 2013.

[113] GSMA. Gsma guidelines on the protection of privacy in the use of mobile phone data for responding to the ebola outbreak. `http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2014/11/GSMA-Guidelines-on-protecting-privacy-in-the-use-of-mobile-phone-data-for-respo October-2014.pdf`, 2014.

[114] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[115] Melissa Gymrek, Amy L McGuire, David Golan, Eran Halperin, and Yaniv Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[116] Eran Hammer-Lahav. Rfc 5849: The oauth 1.0 protocol. *Internet Engineering Task Force (IETF)*, 2010.

[117] Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft Research Redmond, WA, 2009.

[118] Cesar A Hidalgo and C Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024, 2008.

[119] J.I. Hong and J.A. Landay. An architecture for privacy-sensitive ubiquitous computing. In *Proceedings of the 2nd international conference on Mobile systems, applications, and services*, pages 177–189. ACM, 2004.

[120] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856, 2007.

[121] ITU. Ict facts and figures. `http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFacts-Figures2013-e.pdf`.

[122] Clippinger J. in rules for growth: Promoting innovation and growth through legal reform. 2010.

[123] Oliver P John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[124] A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2):618–644, 2007.

[125] Judy Kay and Bob Kummerfeld. Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4):24, 2012.

[126] M. Kearns, J. Tan, and J. Wortman. Privacy-preserving belief propagation and sampling. *Advances in Neural Information Processing Systems*, 20, 2007.

[127] Coco Krumme, Alejandro Llorente, Manuel Cebrian, Esteban Moro, et al. The predictability of consumer visitation patterns. *Scientific reports*, 3, 2013.

[128] MIT Human Dynamics Lab. Reality commons. `http://realitycommons.media.mit.edu/`. Accessed: 2015-01-30.

[129] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[130] S. Lederer, J. Mankoff, and A.K. Dey. Who wants to know what when? privacy preference determinants in ubiquitous computing. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 724–725. ACM, 2003.

[131] Senator Lee. Senate bill 548: Maryland personal information protection act - revisions. `http://mgaleg.maryland.gov/2015RS/bills/sb/sb0548f.pdf`, 2015.

[132] Emmanuel Letouze and Johannes Jutting. *Official Statistics, Big Data and Human Development: Towards a New Conceptual and Operational Approach.* 2014.

[133] N. Li, T. Li, and S. Venkatasubramanian. Closeness: A new privacy measure for data publishing. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):943–956, 2010.

[134] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[135] Antonio Lima, Manlio De Domenico, Veljko Pejovic, and Mirco Musolesi. Exploiting cellular data for disease containment and information campaigns strategies in country-wide epidemics. *arXiv preprint arXiv:1306.4534*, 2013.

[136] Edmond Locard. *Traité de criminalistique.* Joannès Desvigne et ses fils, 1931.

[137] Richard Lynn and Terence Martin. Gender differences in extraversion, neuroticism, and psychoticism in 37 nations. *The Journal of social psychology*, 137(3):369–373, 1997.

[138] Robert C MacCallum, Shaobo Zhang, Kristopher J Preacher, and Derek D Rucker. On the practice of dichotomization of quantitative variables. *Psychological methods*, 7(1):19, 2002.

[139] Carolyn MacCann, Angela Lee Duckworth, and Richard D Roberts. Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences*, 19(4):451–458, 2009.

[140] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. *l*-Diversity: privacy beyond *k*-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)*, page 24, 2006.

[141] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[142] Katrina Manson. From oil painter to the c-suite. 2013.

[143] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

[144] S. Mascetti, D. Freni, C. Bettini, X.S. Wang, and S. Jajodia. Privacy in geo-social networks: proximity notification with untrusted service providers and curious buddies. *The VLDB Journal—The International Journal on Very Large Data Bases*, 20(4):541–566, 2011.

[145] Jonathan Mayer and Patrick Mutchler. Metaphone: the sensitivity of telephone metadata. *Web Policy*, 2014.

[146] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Personality: critical concepts in psychology*, 60:295, 1998.

[147] Marcia McNutt. Journals unite for reproducibility. *Science*, 346(6210):679–679, 2014.

[148] Sandro Meloni, Nicola Perra, Alex Arenas, Sergio Gómez, Yamir Moreno, and Alessandro Vespignani. Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific reports*, 1, 2011.

[149] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM, 2004.

[150] Darakhshan J Mir, Sibren Isaacman, Ramón Cáceres, Margaret Martonosi, and Rebecca N Wright. Dp-where: Differentially private modeling of human mobility. In *Big Data, 2013 IEEE International Conference on*, pages 580–588. IEEE, 2013.

[151] A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel. Movement data anonymity through generalization. *Transactions on Data Privacy*, 3(2):91–121, 2010.

[152] L. Mui, M. Mohtashemi, and A. Halberstadt. A computational model of trust and reputation. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, pages 2431–2439. IEEE, 2002.

[153] M. Mun, S. Hao, N. Mishra, K. Shilton, J. Burke, D. Estrin, M. Hansen, and R. Govindan. Personal data vaults: a locus of control for personal data streams. In *Proceedings of the 6th International Conference*, page 17. ACM, 2010.

[154] Arvind Narayanan. No silver bullet: De-identification still doesn't work. `https://freedom-to-tinker.com/blog/randomwalker/no-silver-bullet-de-identification-still-doesnt-work/`.

[155] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[156] Mark Newman. *Networks: an introduction.* Oxford University Press, 2010.

[157] BBC News. Ebola: Can big data analytics help contain its spread? `http://www.bbc.com/news/business-29617831`.

[158] Nielsen. How smartphones are changing consumers' daily routines around the globe. 2014.

[159] US Department of Health and Human Services. Privacy of individually identifiable health information. 45 C.F.R. 164.514.

[160] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[161] C. Orlandi. Is multiparty computation any good in practice? In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5848–5851. IEEE, 2011.

[162] Venkata N Padmanabhan, Ramachandran Ramjee, and Prashanth Mohan. System for sensing road and traffic conditions, April 16 2013. US Patent 8,423,255.

[163] J. Palfrey and J. Zittrain. Better data for a better internet. *Science*, 334(6060):1210–1211, 2011.

[164] Vijay Pandurangan. On taxis and rainbows, lessons from nyc's improperly anonymized taxi logs. https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1.

[165] Article 29 Data Protection Working Party. Opinion 05/2014 on anonymisation techniques. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, 2014.

[166] Alex Pentland. Society's nervous system: building effective government, energy, and public health systems. 2011.

[167] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60. ACM, 2008.

[168] Rubin C. Azarbayejani A. Place, S. and J. Feast. Evaluation of trust framework sharing and privacy concerns. In *Technical Report 030113*. Cogito Corporation, March, 2013.

[169] L. Popa, M. Yu, S.Y. Ko, S. Ratnasamy, and I. Stoica. Cloudpolice: taking access control out of the network. In *Proceedings of the Ninth ACM SIGCOMM Workshop on Hot Topics in Networks*, page 7. ACM, 2010.

[170] Raluca Ada Popa, Catherine Redfield, Nickolai Zeldovich, and Hari Balakrishnan. Cryptdb: protecting confidentiality with encrypted query processing. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*, pages 85–100. ACM, 2011.

[171] U.N. Global Pulse. Mapping the risk-utility landscape of mobile phone data for sustainable development & humanitarian action. http://www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Mobile_Data_Privacy_2015.pdf, 2015.

[172] Troy Raeder, Omar Lizardo, David Hachen, and Nitesh V Chawla. Predictors of short-term decay of cell phone contacts in a large scale communication network. *Social Networks*, 33(4):245–257, 2011.

[173] J. Reades. Finite state machines: preserving privacy when data-mining cellular phone networks. *Journal of Urban Technology*, 17(1):29–40, 2010.

[174] Federal Reserve. Federal reserve financial services federal reserve study shows more than three-quarters of noncash payments are now electronic. 2010.

[175] Ira S Rubinstein. Big data: The end of privacy or a new beginning? 2012.

[176] Higginbotham S. For science, big data is the microscope of the 21st century. http://gigaom.com/2011/11/08/for-science-big-data-is-the-microscopeof-the-21st-century/. Accessed: 2015-01-30.

[177] K Schwab, A Marcus, JO Oyola, W Hoffman, and M Luzi. Personal data: The emergence of a new asset class. In *An Initiative of the World Economic Forum*, 2011.

[178] Paul M Schwartz and Daniel J Solove. Reconciling personal information in the united states and european union. *Cal. L. Rev.*, 102:877, 2014.

[179] P.M. Schwartz. Property, privacy, and personal data. *Harv. L. Rev.*, 117:2056, 2003.

[180] Maarten Selfhout, William Burk, Susan Branje, Jaap Denissen, Marcel Van Aken, and Wim Meeus. Emerging late adolescent friendship networks and big five personality traits: A social network approach. *Journal of personality*, 78(2):509–538, 2010.

[181] Erez Shmueli, Tamir Tassa, Raz Wasserstein, Bracha Shapira, and Lior Rokach. Limiting disclosure of sensitive data in sequential releases of databases. *Information Sciences*, 191:98–127, 2012.

[182] Erez Shmueli, Ronen Vaisenberg, Ehud Gudes, and Yuval Elovici. Implementing a database encryption solution, design and implementation issues. *Computers & Security*, 2014.

[183] Stanislav Sobolevsky, Michael Szell, Riccardo Campari, Thomas Couronné, Zbigniew Smoreda, and Carlo Ratti. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PloS one*, 8(12):e81707, 2013.

[184] Ayla Solomon, Raquel Hill, Erick Janssen, Stephanie A Sanders, and Julia R Heiman. Uniqueness and how it impacts privacy in health-related social science datasets. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 523–532. ACM, 2012.

[185] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[186] Yi Song, Daniel Dahlmeier, and Stephane Bressan. Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. *ACM PIR*, 2014.

[187] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 321–330. ACM, 2012.

[188] Kristin Brooke Stecher and Scott Counts. Spontaneous inference of personality traits and effects on memory for online profiles. In *ICWSM*, 2008.

[189] Robert Steinbrook. Personally controlled online health data-the next big thing in medical care? *New England Journal of Medicine*, 358(16):1653, 2008.

[190] Arkadiusz Stopczynski, Riccardo Pietri, Alex Pentland, David Lazer, and Sune Lehmann. Privacy in sensor-driven human data collection: A guide for practitioners. *arXiv preprint arXiv:1403.5299*, 2014.

[191] Pål Sundsøy, Johannes Bjelland, Asif M Iqbal, Yves-Alexandre de Montjoye, et al. Big data-driven marketing: How machine learning outperforms marketers' gut-feeling. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 367–374. Springer, 2014.

[192] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671:1–34, 2000.

[193] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002.

[194] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[195] Latanya Sweeney. Matching known patients to health records in washington state data. *Available at SSRN 2289850*, 2013.

[196] Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, 2010.

[197] Deutsche Telekom. Deutsche telekom. `www.telekom.com/static/-/205808/1/guiding-principles-big-data-si`, 2014.

[198] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment*, 1(1):115–125, 2008.

[199] Scott Thurm and Yukari Iwatani Kane. Your Apps Are Watching You . *The Wall Street Journal*, 2014.

[200] European Union. Directive 95/46/ec, recital 26.

[201] Simine Vazire. Who knows what about a person? the self–other knowledge asymmetry (soka) model. *Journal of personality and social psychology*, 98(2):281, 2010.

[202] Hannu Verkasalo, Carolina López-Nicolás, Francisco J Molina-Castillo, and Harry Bouwman. Analysis of users and non-users of smartphone applications. *Telematics and Informatics*, 27(3):242–255, 2010.

[203] K. Wang and B.C.M. Fung. Anonymizing sequential release. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, pages 414–423, 2006.

[204] R. Want, T. Pering, G. Danneels, M. Kumar, M. Sundar, and J. Light. The personal server: Changing the way we think about ubiquitous computing. *Ubicomp 2002: Ubiquitous Computing*, pages 223–230, 2002.

[205] Samuel D Warren and Louis D Brandeis. The right to privacy. *Harvard law review*, pages 193–220, 1890.

[206] Duncan J Watts. *Everything is obvious:* Once you know the answer.* Crown Business, 2011.

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

[208] Amy Wesolowski, Caroline O Buckee, Linus Bengtsson, Erik Wetter, Xin Lu, and Andrew J Tatem. Commentary: containing the ebola outbreak-the potential and challenge of mobile network data. *PLoS currents*, 6, 2014.

[209] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[210] S. William. *Computer Security: Principles And Practice.* Pearson Education India, 2008.

[211] Matthew J Williams, Roger M Whitaker, and Stuart M Allen. Measuring individual regularity in human visiting patterns. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 117–122. IEEE, 2012.

[212] X. Xiao and Y. Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of Data (SIGMOD '07)*, pages 689–700.

[213] Jane Yakowitz. Tragedy of the data commons. *Harv. JL & Tech.*, 25:1, 2011.

[214] R. Yarovoy, F. Bonchi, L.V.S. Lakshmanan, and W.H. Wang. Anonymizing moving objects: how to hide a mob in a crowd? In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 72–83. ACM, 2009.

[215] Hui Zang and Jean Bolot. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, pages 145–156. ACM, 2011.

[216] G. Zhong, I. Goldberg, and U. Hengartner. Louis, lester and pierre: Three protocols for location privacy. In *Proceedings of the 7th international conference on Privacy enhancing technologies*, pages 62–76. Springer-Verlag, 2007.