Privacy Analytics

Hamed Haddadi
Queen Mary
University of London, UK
hamed@eecs.qmul.ac.uk

Richard Mortier
University of Nottingham, UK
richard.mortier@nottingham.ac.uk

Steven Hand
University of Cambridge, UK
steven.hand@cl.cam.ac.uk

Ian Brown
Oxford Internet Institute, UK
ian.brown@oii.ox.ac.uk

Eiko Yoneki,
Jon Crowcroft
University of Cambridge, UK
eiko.yoneki@cl.cam.ac.uk,
jon.crowcroft@cl.cam.ac.uk

Derek McAuley
University of Nottingham, UK
derek.mcauley@nottingham.ac.uk

ABSTRACT

People everywhere are generating ever-increasing amounts of data, often without being fully aware of who is recording what about them. For example, initiatives such as mandated smart metering, expected to be widely deployed in the UK in the next few years and already attempted in countries such as the Netherlands, will generate vast quantities of detailed, personal data about huge segments of the population. Neither the impact nor the potential of this society-wide data gathering are well understood. Once data is gathered, it will be processed – and society is only now beginning to grapple with the consequences for privacy, both legal and ethical, of these actions, e.g., Brown *et al.* [4]. There is the potential for great harm through, e.g., invasion of privacy; but also the potential for great benefits by using this data to make more efficient use of resources, as well as releasing its vast economic potential [28]. In this editorial we briefly discuss work in this area, the challenges still faced, and some potential avenues for addressing them.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: SOCIAL AND BEHAVIORAL SCIENCES

General Terms

Design, Human Factors, Security

Keywords

Privacy, Aggregation, Mobility, Surveys, Profiling

1. PRIVACY PRESERVING DATA MARKET

People everywhere are generating ever-increasing amounts of data, often without being fully aware of who is recording what about them. Similarly, governments, industries and research organisations increasingly demand public data be made available to them. Unfortunately, releasing large public datasets often has consequences for privacy, both legal and ethical, e.g., Brown *et al.* [4]. We propose constructing a framework – *Privacy Analytics* – enabling querying of such data in such a way as to avoid these consequences by first verifying query code, and then launching it into the user community to perform its measurement tasks, collect verifiable statistics, and finally perform aggregation and fuzzing while remaining within the community.

The Privacy Analytics[1] framework uses the Dataware framework [18] to enable a number of independent queries and measurements across a network of users to be carried out *without* leaking intermediate results and potentially compromising privacy. This framework will allow long-term, detailed and complex queries to be carried out: the output data is limited in entropy and its probabilistic inclusion and aggregation across many cliques and groups of individuals leads to a distributed form of differential privacy [9]. A final stage of data aggregation and statistical dilution is performed before the results are finally presented to the query provider.

Successful deployment of this system necessitates bringing together many disparate areas of research: information theory, sampling theory, distributed systems, measurement and monitoring, differential privacy, mobile computing and data mining. This editorial discusses the fundamental challenges posed by such a system from several viewpoints, including privacy, usability, security and system design. Our aim in designing and building a Privacy Analytics framework is to build in privacy from the ground up, enabling the user to exercise meaningful choice over participation and what personal information they reveal.

Scenarios

To focus discussion we concentrate on four example scenarios where we believe our Privacy Analytics framework will be extremely beneficial:

Streaming media market research.

The modern media industry uses a range of digital media delivery mechanisms including digital broadcast, live streaming and on-demand via the Internet. However, media organisations have only limited insight into the consumption of their media. Organisations such as the BBC have fine-grained data concerning online media consumption, but not in a readily usable form. If dealt with in a privacy preserving manner such data could provide statistics about choice

[1]Analytics is used in the general context of consumer and usage analysis of services and products, rather than the specific tracking by third-parties of website use, e.g., Google Analytics, which has led to a number of *Do Not Track* projects and tools [25].

of content, and manner and time of consumption. These statistics could be gathered to enable analyses across different user categories, grouping users by location, profile and other features. Such information can be used for program timing optimisation, user behavioural analysis and predictive targeted personalised advertising.

Smart energy metering.

In the next few years many, if not most, electricity meters will be replaced by smart meters giving networked access to meter readings. Governments, energy providers and various industrial organisations want to understand consumers' energy consumption in detail and in aggregate, but public acceptance of such techniques will require avoiding risking users' privacy [22]. Users themselves are also interested in accessing such data to monitor their usage, whether at home, at work or travelling. Users' mobile handsets could act as an interface to the Privacy Analytics Framework, enabling privacy preserving access to fine-grained energy consumption data. Without such measures, smart metering programmes risk being derailed by severe public backlash and legal challenges, as happened in the Netherlands [7].

Transport and location privacy.

Users' mobile devices act as sensors, giving information about their environment and behaviour. For example, they can record users' locations giving access to movement patterns and so potentially modes of transport and related energy consumption. As well as providing information of value to the user, this data could provide insight for public transport operators, whether roads management, rail operators, bus companies or government agencies, while monitoring queries which may not be approved by users.[2] For example, long-term commuter trends would inform capacity planning; mass crowd movements at events would aid public safety management; and detection of unexpected traffic jams can provide evidence of significant localised effects such as crashes.

Market research and advertising.

Companies in many industries, from retail to insurance, wish to understand the trends in individuals' shopping habits, expenditures and incentives, and privacy issues are the biggest barrier to such studies. It is likely that individuals would be happy to take part in such surveys in return for monetary benefits if their information is kept with privately, or in their own Personal Container [20]. The information gained can also be used by privacy-preserving targeted advertising systems such as MobiAd [14]. Within the Privacy Analytics framework we aim to understand the feasibility of a system whereby both parties benefit from market research queries.

In the rest of this paper we discuss a number of existing research methods and future trends within the UK and international *Digital Economy* research programme, and then present some of the main challenges faced when dealing with personal data collection and analysis in these scenarios. We note that many of these challenges arise, or will arise, more widely than the scenarios above, in industries such as healthcare, banking, finance and retail.

[2]`http://news.cnet.com/8301-30685_3-20058228-264.html`

2. RELATED WORK

If individuals are prepared to trust a third-party with access to detailed records of their activities, that third-party could choose to execute queries with statistical controls, returning aggregate data that reduces the information leaked to the querier about the individual. For example, in health studies, Loukides *et al.* propose an algorithm that protects patients' personal information while preserving the data's utility in large-scale medical studies [17]. This is done by broadening the category groups that patients fall into, and hence decreasing the risk of identifying individuals, at the cost of less accurate medical correlation between genes and medical conditions. Without very careful privacy controls, the amount of information present in such statistics can easily be used to identify the individuals concerned [27].

Effective privacy controls must be layered on top of effective security. This remains true when user data is stored on personal devices such as smartphones, increasingly the targets of malware.[3] If more sensitive data is to be stored about users, we need a better understanding of how appropriate security protection can be provided through, for example, the trustworthy hardware components being explored in projects such as Webinos.[4] Users also need more effective tools for deciding whether they trust specific software with access to their personal data, which could be based on distributed recommendation systems such as Convergence.[5]

Users' trust is critically dependent on their confidence in full control over data collection and use [3]. Twenty-page legalistic privacy policies that an individual must click to accept before using software do not provide this confidence; and nor do default privacy settings that open up user data to unexpected purposes and recipients. Explicit, easily reversible, opt-in usage is the most meaningful way to provide effective user control, and will also give better compliance with the range of data protection laws increasingly prevalent around the world [12]. We also must consider the potential for compelled access to data by employers, courts, government agencies, and other powerful institutions, as well as access by nefarious parties such as hackers and criminals. Minimisation of personal data storage remains important even when it is kept on devices under the user's effective control [5].

There have been recent studies on analysing network traces using differential privacy [19], and on accessing databases while respecting privacy [15], but there has not yet been an operational system that also helps utilise and expose statistics and trends on information for outsiders. Rieffel *et al.* [24] propose cryptographic, hierarchical access to data for processing aggregate statistics without decrypting personal data. However this method still requires collection of individual data items and key management.

The FluPhone project [29] targeted tackling 'flu-like symptoms, following the perceived threat of bird 'flu in our society a couple of years ago. Human proximity information is collected from the general population using phones with Bluetooth communication, to build time dependent contact networks. The project also included a 'virtual disease' ex-

[3]`http://news.cnet.com/8301-1009_3-57328575-83/androids-a-malware-magnet-says-mcafee/`

[4]`http://webinos.org/about-webinos/`

[5]`http://convergence.io/`

periment, where a specific model of disease is spread through the proximity based communication upon encountering of two devices. The spread of different stages of the disease was then mapped across the locality of the study and fed back to the user. The collected data is valuable, but currently its analysis is limited due to lack of a clear understanding as to how much privacy could be leaked. The research outcome will empower medics to explore research using real-world data and benefit for users to entrust more personally identifiable information. In Privacy Analytics we will also use mobile based agents and Crowd Computing concept by Murray *et al.* [21] to achieve the in-community aggregation goals.

3. CHALLENGES

There are a number of challenges faced by research and industry when it comes to using personal information. The needs of the individual have been completely ignored in the rush to large-scale online data mining. We identify the following particular challenges that we will address within the Privacy Analytics framework.

- A wide range of mechanisms exist to provide users with degrees of control over information based on models of privacy, as briefly discussed above. Unfortunately, much of this work is based on theoretical work, e.g., [16], with little in the way of ground-truth concerning the details of users' perception of the value of private information. We need to understand altruistic and selfish, i.e., induced by monetary reward, behaviours in participation selection.

- Many current projects, e.g., the Locker Project,[6] are building personal information management systems that bring an individual's data together for them to manipulate and manage. However, the trade-offs between security, privacy and usability of such personal profiling and information gathering systems are not well understood – what vulnerabilities are introduced by centralising your data, and what opportunities are created as a result? Such analysis would also directly inform governments' use of consumer data for expenditure advice [26].

- Privacy concerns arise in the Internet in a wide range of contexts, using a wide range of technologies and devices. These contexts need to be better understood, and definitions formalised, to enable quantification of the compromises available to users between privacy leakage and the benefits available to both users and service providers. Given the rapid expansion in mobile, and particularly smartphone usage, this is particularly important in mobile.

4. OUR APPROACH

Detailed profiling and interest mining has been the basis for operation of online retailers and services such as as Facebook, Google and Amazon. However, it is also well known that such profiling exposes the user to privacy leakage, even when these communities are anonymous [8]. Privacy-preserving advertising methods [13, 14] aim to eliminate centralised user profiling and keep the user profile at the end host. We wish to explore the possibility of using these end host profiles for carrying large scale surveys and market research statistics. Such systems would require building up individual's demographic information and filtering accordingly. In this section we explore some of the methods of dealing with these privacy challenges.

4.1 Privacy by Data Aggregation

With any information collection system, there is a trade-off between amount of information collected versus the impact on individuals' privacy. A highly targeted survey can lead to a high level of privacy leakage, while a poorly anonymised and distorted dataset can lead to results which are so divergent from reality that they are of no use at all. Lower bounds of query outcome must be investigated alongside the sampling theory implications. As a few examples, data dilution/fuzzing at various stages can be achieved by:

Distribution building. In this scenario, a distribution of result sets is the output. The number of bins in a histogram can depend on the sensitivity of the data. Alternatively a correlation coefficient or probability can be presented as the query output.

Sanitisation. A certain amount of noise can be added to the data in order to decrease its accuracy. A final answer with an error term can be output. Some work has already done on the theory underlying this approach [10][2].

Crowd anonymisation. When any query is carried out, any individual's data is taken into account probabilistically. While this preserves the general distribution shape and statistics, it will prevent an individual's data to be identifiable even if outcome is reverse engineered.

Coarse-grained data. This is particularly relevant for location-specific surveys and queries. The data can be coarsened to alleviate location-sensitive data using techniques such as *k-Anonymity* [6]. For example local data collected at different parts of a town can be aggregated if the number of subjects/participants is below the minimum bound imposed by the query's privacy implications.

4.2 Profiling the Community

User communities evolve over time so it is also important that a group/clique is defined in the right context both for aggregation and for ad targeting. We aim to use different real time community detection methods on users' profiles (social network data) and handsets (location/contact data) to perform this. However in this context, accuracy is not critical and this method is also time-insensitive, within established bounds, to aid privacy.

As part of this work we wish to establish the community dynamics in social networks, infer contextual cliques and compare their characteristics with those inferred from human-contact data already collected. By correlating these two forms of networks we wish to understand the feasibility of multi-layer aggregation across different resolution of communities. This will aid with aggregation strategies as shown in Figure 1.

[6] `http://lockerproject.org/`

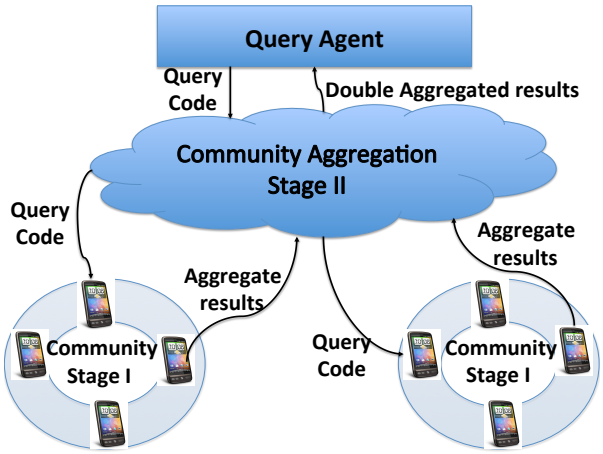

Figure 1: Query and result aggregation overview.

4.3 User Incentives

We must understand the incentives users have concerning privacy of their data: are people prepared to sacrifice privacy (sex, age, location, etc.) for gain, either voluntarily or monetary in the trade-offs that most Internet users are willing to make with some personal data. If users' details can be used to perform market research, or medical surveys or for targeted advertising, while preserving privacy, will they participate? We must understand and formalise the definition of privacy in different contexts for mobile and Internet users, and provide ways to quantitatively measure the compromise between privacy leakage and benefit to users and service providers.

5. THE FUTURE OF PERSONAL DATA

The overarching aim of the Privacy Analytics Framework is to dis-intermediate the cloud's advertising-based revenue model, for two reasons:

- Providing user privacy, from both advertisers/market researchers *and* cloud service providers, prevents the former from unnecessary invasion and the latter from the temptation to examine personally identifiable information.
- Freeing up the cloud advertising market so that, rather than simply having all of the revenue from traditional media advertising (TV, radio, print) going to the cloud – which is to say Google and Facebook currently – we enable more widespread *socially acceptable* use of the exquisitely accurate targeting and collection opportunities afforded.

The latter reason deserves a little more discussion. We do not refer only to existing models of click-through tracking, but to the more detailed information potentially available. For example, *actual* sales of goods and services tells not just whether or not an advert got a user's attention but what that attention was actually worth in increased sales. Obviously individual retailers such as Amazon already have this, but pure-play advertisers like Google and Facebook do not know what was subsequently acted upon, notwithstanding the search, mail, social and other properties they own. Thus clients pay to get their ads higher up in ranking systems, and neither advertisers nor cloud providers can truly connect the price paid with the resulting profit generated.

Thus, by deploying Privacy Analytics, we face the customer and entice them to *increase* their personal contextual footprints – the digital data available concerning their online and offline lives. This enticement arises because we guarantee to protect this data about them; while also facing the goods and service providers and offering them the ability to determine both the effectiveness (and thus *price*) of an advert and the demographics of its effectiveness, without compromising said privacy. This is a win-win-lose-win scenario as the users, advertisers, and retail/wholesale goods/service providers all win, but Google/Facebook may lose some revenue since it takes the near-monopoly world of online targeted adverts, and turns it into a competitive market where profits should be marginal.

To build such a marketplace raises a further important challenge concerning users' awareness of the value of their data. Addressing this requires a major study: will users take part in such a scheme altruistically, or must they be incentivised? How? Despite a number of recent systematic approaches to selling private information [23, 11], it is extremely difficult to measure individuals' evaluation of their privacy as their perception changes under different circumstances [1]. As part of Privacy Analytics, we will devise a methodology for understanding the users' perception of privacy and its monetary value

Individual privacy rights seem to have been of secondary concern in the personal data gold rush of recent technology development. This is partly because privacy regulators have had difficulty keeping up with the rate of technology change, and partly because the new technology had to find a way to make value from giving away content otherwise the revenue stream for new media (music, film etc) would have dried up and those business sectors would have just died completely. The goal of Privacy Analytics and similar projects is to rebalance these rights without disrupting the new business models. However, there are no definitions of absolute or eternal privacy, so technology can only be part of the solution and a larger effort by standards agencies, government organisations and regulators is required to enable true control over users' privacy.

In general, the Privacy Analytics Framework will enable growth of a new ecology of social and economic applications based around large-scale processing of personal data. By providing the technical means for this while understanding, quantifying and respecting the privacy concerns of users, commercial and social organisations will have access to much larger, much richer data sources than currently possible.

Acknowledgments

This work was funded by the RCUK Horizon Digital Economy Research Hub grant, EP/G065802/1. We acknowledge feedback from Claude Castelluccia and anonymous reviewers.

6. REFERENCES

[1] A. Acquisti, L. John, and G. Loewenstein. What is Privacy worth? In *Proceedings of the Twenty First Workshop on Information Systems and Economics (WISE)*, Dec. 2009.

[2] G. Ács and C. Castelluccia. I have a dream!: differentially private smart metering. In *Proceedings of the 13th international conference on Information hiding*, IH'11, pages 118–132, Berlin, Heidelberg, 2011. Springer-Verlag.

[3] A. Adams and M. A. Sasse. *Privacy in Multimedia Communications: Protecting Users, Not Just Data*. 2001.

[4] I. Brown, L. Brown, and D. Korff. Using NHS patient data for research without consent. *Law, Innovation and Technology*, 2(2):219–258, Dec. 2010.

[5] I. Brown and B. Laurie. Security against compelled disclosure. In *In Computer Security Applications 16th Annual Conference (ACSAC '00). IEEE*, pages 2–10, 2000.

[6] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. k-Anonymity. In T. Yu and S. Jajodia, editors, *Secure Data Management in Decentralized Systems*. Springer-Verlag, 2007.

[7] C. Cuijpers. No to mandatory smart metering does not equal privacy! In *Tilburg Institute for Law, Technology, and Society*, 2009.

[8] C. Díaz, C. Troncoso, and A. Serjantov. On the impact of social network profiling on anonymity. In *Privacy Enhancing Technologies*, pages 44–62, 2008.

[9] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin / Heidelberg, 2006.

[10] C. Dwork. A firm foundation for private data analysis. *Commun. ACM*, 54(1):86–95, 2011.

[11] A. Ghosh and A. Roth. Selling privacy at auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, pages 199–208, New York, NY, USA, 2011. ACM.

[12] G. Greenleaf. *Global Data Privacy in a Networked World*. 2011.

[13] S. Guha, A. Reznichenko, K. Tang, H. Haddadi, and P. Francis. Serving ads from localhost for performance, privacy, and profit. In *Eighth ACM Workshop on Hot Topics in Networks (HotNets-VIII)*, New York City, NY, 2009.

[14] H. Haddadi, P. Hui, and I. Brown. Mobiad: private and scalable mobile advertising. In *Proceedings of the fifth ACM International Workshop on Mobility in the Evolving Internet Architecture*, MobiArch '10, pages 33–38, New York, NY, USA, 2010. ACM.

[15] C. M. Johnson and T. W. A. Grandison. Compliance with data protection laws using hippocratic database active enforcement and auditing. *IBM Systems Journal*, 46(2):255 –264, 2007.

[16] J. Kleinberg, C. H. Papadimitriou, and P. Raghavan. On the value of private information. In *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '01, pages 249–257, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[17] G. Loukides, A. Gkoulalas-Divanis, and B. Malin. Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107(17):7898–7903, April 2010.

[18] D. McAuley, R. Mortier, and J. Goulding. The Dataware Manifesto. In *Proceedings of the 3rd IEEE International Conference on Communication Systems and Networks (COMSNETS)*, Bangalore, India, January 2011. Invited paper.

[19] F. McSherry and R. Mahajan. Differentially-private network trace analysis. *SIGCOMM Comput. Commun. Rev.*, 40:123–134, August 2010.

[20] R. Mortier, C. Greenhalgh, D. McAuley, A. Spence, A. Madhavapeddy, J. Crowcroft, and S. Hand. The Personal Container, or Your Life in Bits. In *Proceedings of Digital Futures*, October 2010.

[21] D. G. Murray, E. Yoneki, J. Crowcroft, and S. Hand. The case for crowd computing. In *Proceedings of the Second ACM SIGCOMM workshop on Networking, Systems, and Applications on Mobile Handhelds*, MobiHeld '10, pages 39–44, New York, NY, USA, 2010. ACM.

[22] A. Rial and G. Danezis. Privacy-preserving smart metering. In *Proceedings of the 10th annual ACM Workshop on Privacy in the Electronic Society*, WPES '11, pages 49–60, New York, NY, USA, Oct. 2011. ACM.

[23] C. Riederer, V. Erramilli, A. Chaintreau, P. Rodriguez, , and B. Krishnamurthy. For Sale : Your Data By : You. In *Proceedings of the ACM/HOTNETS*, 2011.

[24] E. G. Rieffel, J. T. Biehl, W. van Melle, and A. J. Lee. Secured histories: computing group statistics on encrypted data while preserving individual privacy. *CoRR*, abs/1012.2152, 2010.

[25] O. Tene and J. Polonetsky. To track or 'do not track': Advancing transparency and individual control in online behavioral advertising. In *SSRN*, 2011.

[26] UK Cabinet Office. Better Choices: Better Deals. `http://www.cabinetoffice.gov.uk/resource-library/better-choices-better-deals/`, April 2011.

[27] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, CCS '09, pages 534–544, New York, NY, USA, 2009. ACM.

[28] World Economic Forum and Bain, eds. Personal data: The emergence of a new asset class. `http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf`, Jan. 2011.

[29] E. Yoneki and J. Crowcroft. Epimap: Towards quantifying contact networks and modelling the spread of infections in developing countries. In *AWCR*, 2011.