On the Trusted Use of Large-Scale Personal Data

Yves-Alexandre de Montjoye* #1, Samuel S. Wang *2, Alex (Sandy) Pentland #3
#The Media Laboratory
*Decentralized Information Group, CSAIL
Massachusetts Institute of Technology
Cambridge, MA, USA
[1]yva@mit.edu, [2]samuelsw@csail.mit.edu, [3]pentland@mit.edu

Abstract

Large-scale personal data has become the new oil of the Internet. However, personal data tend to be monopolized and siloed by online services which not only impedes beneficial uses but also prevents users from managing the risks associated with their data. 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 has yet to be realized and tested. We introduce openPDS, an implementation of a Personal Data Store, which follows the recommendations of the WEF, the US NSTIC and the US Consumer Privacy Bill of Rights. openPDS allows users to collect, store, and give fine-grained access to their data in the cloud. openPDS also protects users' privacy by only sharing anonymous answers, not raw data. Indeed, a mechanism to install third-parties applications on a user's PDS allows for the sensitive part of the data processing to take place within the PDS. openPDS can also engage in privacy-preserving group computations to aggregate data across users without the need to share sensitive data with an intermediate entity.

1 Motivation

Personal Data has become the new oil of the Internet [1], 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 thus 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 do 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 installing third-party applications, the system ensures that most processing of sensitive personal data takes place

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

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 [2], the US National Strategy for Trust Identities in Cyberspace (NSTIC) [3], the Department of Commerce Green Paper, and the Office of the Presidents International Strategy for Cyberspace [4]. 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 [5]. 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.

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 [1]. 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.

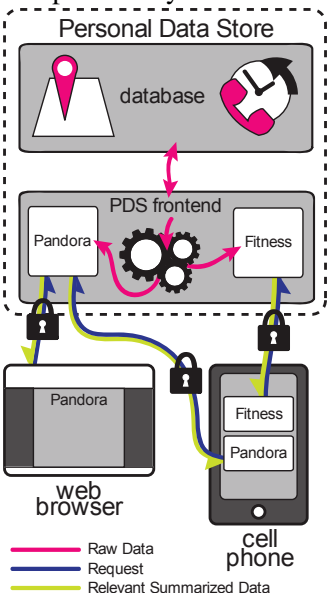
Thanks to the policy requirement of data portability, a PDS-based data market is likely to be economically efficient, as the system removes barriers to entry for new businesses. It allows the more innovative companies to provide better data-powered services. The services chosen by the user will have 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 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 [See Fig. 1].

Figure 1: openPDS system's architecture.



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.

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 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.

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.

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.

References

[1] Personal Data: The Emergence of a New Asset Class, http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf.

[2] US Consumer Privacy Bill of Rights, http://www.whitehouse.gov/sites/default/files/privacy-final.pdf

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

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

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

[6] 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