Private Data Exchange

Ashwin Machanavajjhala, Yahoo! Research, Santa Clara, CA, USA
mvnak@yahoo-inc.com

1 Introduction

The personal information that users generate on the Web, social networks and that is continuously tracked via mobile phones can enable new science and can be used to provide valuable services to users. However, most of this information is tracked and stored in vaults by private industries. Health data is vaulted away in hospital and insurance databases, while search data is exclusively owned by Web search companies. This leads to many problems – 1) there is no mechanism for individual to own and get remunerated for their valuable data, 2) there is no ways users can track what is being done with their data and 3) scientific endeavors that might want to join e.g., medical with Web data are not possible since the data is not available in one place.

I envision a future where users retain ownership of their personal information and store it in one secure location on the cloud. I call this the Private Data Exchange (PDX). Not only can users now track all their data, they can allow third parties to write applications on top of this data, and can "sell" their data to either companies in health care, social networks, or to research e.g., from the Census in return for either monetary compensation or services. I believe that with the advances in cloud technology, privacy mechanisms and our understanding of markets, this vision can indeed become a reality.

2 Private Data Exchange

A *Private-Data Exchange*, or PDX, is a system where individuals store all their personal information in one place on the cloud, and registered applications and service providers, e.g., insurance companies, advertisers, or researchers, can negotiate with the individuals to use their personal information in return for monetary or service-based incentives. Such a system would help users manage and monitor their personal data in a single place, track breaches of privacy, and allow them to be remunerated for sharing their information. It would also help build novel services by integrating disparate kinds of information. PDX would have the following key components: (a) a secure cloud-based storage infrastructure, where users can store and retrieve their personal information, (b) mechanisms for application developers to write queries (including long-standing continuous queries) over user data, (c) a query optimizer, which tracks the requests for data from various service providers, and generates the right answer based on prior queries being answered and the user's privacy settings, (d) a privacy negotiator, that allows users to trade-off utility for privacy with service providers, and finally (e) an auditor, which ensures that personal information is accessed according to correctly negotiated terms of use. Building such a system has many interesting research challenges. While there is ongoing work on some of the research questions, like secure cloud storage [1], and privacy auditing [2], I outline below a selected set of research questions related to privacy and big-data management that would provide the necessary set of tools and concepts to realize this vision.

3 Research Questions

• Privacy Fundamentals Research

In order to understand the limits of the envisioned Private Data Exchange, the following key fundamental questions in privacy must be answered. Recent research [3, 4, 5] has shown that state-of-the-art privacy definitions assume a "worst case" adversary that are either unrealistic, leading to poor utility, or do not represent the worst case for data where individuals are correlated, leading to privacy breaches. Hence, current privacy mechanisms are not applicable to realistic data arising on the Web and social networks, and in continuous monitoring applications. Moreover, results like the No Free Lunch Theorem [3] show that there is no single privacy definition for all kinds of data. For instance, it was shown that existing privacy mechanisms (like differential privacy) only work for data without correlations, and either leak information

or provide no utility in social network related data sharing. Hence, finding the right privacy definitions and mechanisms for social networks and continuous data collection is an important open problem. More ambitiously one can think of building a customizable privacy compiler tool that can generate domain specific mechanisms with formal privacy guarantees.

Next, an obstacle for the adoption of current state-of-the-art privacy definitions is that it is hard for an individual or an application developer to understand what sensitive information maybe breached to a realistic adversary. It would be nice to build tools that, given a proposed mechanism for data sharing, automatically identifies examples of possible privacy breaches. This tool can then elicit user feedback as to which breaches are tolerable and which are not, and accordingly can tune the privacy mechanism. Finally, integrating utility-theoretic notions of privacy into the prevalent mathematical worst-case notions would be required to better understand how individuals trade-off utility for privacy.

• Big-Data Management Challenges

The envisioned PDX system brings forth many data management challenges, including those important in other settings as well, in building and managing systems that can efficiently and securely store varied forms of personal information. First, feed-following [6] is becoming a very important pattern of real-time data access. For instance, a user in a social networking, an emergency responder, or a health provider-facing application may want to follow the "feed" of information generated by another user. Queries are then answered on the pertinent set of feeds – e.g., latest k urls tagged by a user's friend on a social network, or the set of users driving at speed $> 85mph$. One can envision queries that need probabilistic inference over an individual's data, e.g., the set of individuals with $P[\text{heart attack } > .8]$ based on a model over an individual's medical history and physical activity. At the same time, unintended recipients must not learn anything about the user. Some key research questions in this problem are outlined in [6].

A second interesting problem is that of view materialization on the cloud for applications that continuously monitor aggregate statistics over a population of individuals. As the size of raw data increases, it is important to intelligently materialize views over the raw data. This is all the more critical if the access must be privacy preserving – every time personal information is accessed, there is a potential for further information disclosure, thus motivating the need for privacy-aware views.

• Novel Applications

Today, individuals share immense amounts of information about themselves online. These include unstructured posts on social networking sites, and activity tracked via continuous monitoring mobile applications. Integrating such information can provide valuable cues about population demographics (with applications to the Census), and about individual and population health (e.g. Google Flu). One concrete research problem is reducing the costs and effort of performing a Census by utilizing the unstructured information on the Web and social networks, as well as employing crowd-sourcing techniques where the other sources do not have sufficient coverage.

References

[1] S. Bajaj and R. Sion. Trusteddb: A trusted hardware based outsourced database engine. *PVLDB*, 4(12):1359–1362, 2011.

[2] D. Garg, L. Jia, and A. Datta. Policy auditing over incomplete logs: Theory, implementation and applications. In *ACM Conference on Computer and Communications Security*, October 2011.

[3] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 international conference on Management of data*, SIGMOD '11, pages 193–204, New York, NY, USA, 2011. ACM.

[4] A. Machanavajjhala, J. Gehrke, and M. Götz. Data publishing against realistic adversaries. *Proc. VLDB Endow.*, 2:790–801, August 2009.

[5] A. Machanavajjhala, A. Korolova, and A. D. Sarma. Personalized social recommendations: accurate or private. *Proc. VLDB Endow.*, 4:440–450, April 2011.

[6] A. Silberstein, A. Machanavajjhala, and R. Ramakrishnan. Feed following: the big data challenge in social applications. In *Databases and Social Networks*, DBSocial '11, pages 1–6, New York, NY, USA, 2011. ACM.