Diffix: GDPR-level Anonymity with High Utility Analytics

Paul Francis
Max Planck Institute for Software Systems
Kaiserslautern Germany
francis@mpi-sws.org

1 Extended Abstract

Two core principles of responsible personal data management are the ability to inform individuals about how their personal data is used, and to allow individuals to control how their personal data is used. These principles are very hard to achieve in practice, both for technical reasons and for a variety of social reasons (the difficulty of understanding risk or the inherent conflict between service providers and users). It stands to reason, then, that any time some service requirement can be satisfied without resorting to personal data, this should be exploited to the full.

Personal data collected by services is often used for analytic purposes. The analysis can benefit the service (providing an understanding of its own operations or customer base), can benefit third parties including society at large, and can directly benefit the user. An example of the latter would be a health and diet application that allows its users to compare themselves with other users, for instance to learn how a change in diet or exercise might affect them.

Many analyses require results only about aggregate data, not individual personal data. Such analyses can in principle be done over anonymized data. If user data is strongly anonymized in the GDPR sense, then it is by law and in fact not personal data [1]. Personal data management systems [2] should therefore exploit the benefits of anonymization whenever and wherever possible.

In the past, data anonymization has been extremely difficult for a variety of reasons. There have been no general-purpose, easy-to-use tools for anonymization. Rather, data anonymization has required a deep understanding of a set of complex anonymization mechanisms (data swapping, aggregation, rounding, masking, noise adding) and how they affect the expected analysis. Thus data anonymization has been complex, difficult, time consuming, and unfortunately prone to error. As a result, data shared for analytic purposes is most often pseudonymized rather than anonymized. Pseudonymization is the process of removing Personally Identifying Information like names and addresses, but otherwise leaving the data intact. Indeed pseudonymization is used often enough (and confused with anonymization often enough) that the EU Working Party 29 opinion on anonymization [**?**]evotes substantial space to discussing the issue. The GDPR considers merely pseudonymized data to be personal data.

This presentation describes Diffix [3], a new approach to database anonymization that offers a substantially better utility/privacy trade-off than existing approaches including K-anonymity or Differential Privacy. Diffix acts as an SQL proxy between the analyst and an unmodified live database. Diffix adds a minimal amount of noise to answers—Gaussian with a standard deviation of only two for counting queries—and places no limit on the number of queries an analyst may make. Diffix works with any type of data, and configuration is simple and data-independent: the administrator does not need to consider the identifiability or sensitivity of the data itself. In short, a service that gathers personal data can in most cases use Diffix for aggregate analytics as it would any database.

Diffix was developed through a research partnership between the Max Planck Institute for Software Systems and the startup Aircloak. While Diffix has received a favorable evaluation by the French national data protection authority CNIL, it would be incorrect to say that CNIL (or anyone) has 100% confidence in its anonymity properties. Exposure and critical evaluation from the broader research community is still required.

In: M. Sjöberg, Y. Kortesniemi, H. Honko, T. Lehtiniemi (eds.): Proceedings of the MyData 2017 Workshop on Technical Issues and Approaches in Personal Data Management, Tallinn, Estonia, 30-08-2017, published at http://ceur-ws.org

In this presentation, we will give an overview of how Diffix works, focusing on the applicability and limits of the technology, and a brief demo or Aircloak's implementation of Diffix. We hope to engage the workshop in a discussion of where Diffix can and cannot be used in existing or planned personal data management systems, as well as the liabilities of anonymization more generally.

References

[1] Waltraut Kotschy The new General Data Protection Regulation - Is there sufficient pay-off for taking the trouble to anonymize or pseudonymize data? https://fpf.org/wp-content/uploads/2016/11/Kotschy-paper-on-pseudonymisation.pdf, Nov. 2016.

[2] 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.

[3] Paul Francis, Sebastian Probst-Eide, Reinhard Munz Diffix: High-Utility Database Anonymization Annual Privacy Forum APF 2017, Vienna, June 2017