

CASE STUDIES OF DE-IDENTIFYING BIG DATA

KHALED EL EMAM

PRIVACY ANALYTICS INC.

KELEMAM@PRIVACY-ANALYTICS.COM

Under existing legal frameworks, de-identification is one of the few tools available to allow the using and disclosing of individual consumer or patient level data for secondary purposes. There have been concerns expressed about the ability to de-identify different types of big data, such as financial transaction data, car ride data, and device data. This is partially driven by public examples of claimed re-identification attacks, and by the dearth of case studies illustrating how to de-identify this kind of data successfully and in a defensible manner.

De-identification methods for large and complex data sets have been in use in the health sector for more than a decade. These methods account for multiple layers of controls that need to be deployed, such as security, privacy, and contractual controls, in addition to perturbations to the data itself. They also are quantitative in that they utilize risk estimation models to assess the probability of a successful re-identification under different contexts. In such cases de-identification is viewed as a risk management exercise. More recently, these risk-based methods have been incorporated into standards and best practice guidelines around the globe.

Over the last two years we have been involved in the de-identification of very large financial transaction data, device data, and car ride data when that data has been used and disclosed for secondary purposes. This de-identification process followed an adapted version of the methodology that has been applied in health settings. The purposes of this presentation are to:

- Describe a risk-based de-identification methodology that has been used to de-identify different types of large data sets.
- Use real case studies covering billions of records to illustrate how this methodology has been applied.
- Share lessons learned from these experiences about the nature of this data, the challenges, and how to achieve defensible risk levels.

The main conclusion is that current standards-based methods can be used to de-identify big data. However, each data set type has its own characteristics that would have an impact on the de-identification assumptions and strategy.