CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 3

29-31 October 2019, the Hague, the Netherlands 18 October 2019

Privacy, confidentiality, disclosure: What is the difference

Krish Muralidhar (University of Oklahoma) and Rathindra Sarathy (Oklahoma State University) (United States of America)

krishm@ou.edu

Abstract and Paper

There are many definitions of privacy, security, and disclosure when it comes to protecting data. We evaluate the common characteristics, differences and relative strengths of these definitions. We explore options for a composite measure that incorporates many of these definitions.

Privacy, confidentiality, disclosure: What is the difference?

Krish Muralidhar* and Rathindra Sarathy**

* Price College of Business, University of Oklahoma, Norman, OK 73072, USA (krishm@ou.edu)
** Spears College of Business, Oklahoma State University, Stillwater OK 74078
(rathin.sarathy@okstate.edu)

Abstract: The issue of statistical disclosure when releasing sensitive data for analysis has received considerable attention in the literature over the last 50 years. The traditional approach to disclosure limitation is to consider alternative techniques to specific data sets to evaluate the trade-off between the extent to which the technique prevents disclosure and the impact of the technique on the usefulness of the data. We refer to these as Statistical Disclosure Limitation (SDL) techniques. A wide variety of approaches have been proposed in the literature.

In recent years, there has been a different philosophy towards the issue of preventing disclosure, focusing on privacy. The first such approach was k-anonymity and the more recent being differential privacy. Also referred to as a "privacy first" philosophy, these approaches attempt to minimize the ability of an adversary to identify an individual from responses to queries and/or data released for analysis purposes. Unlike traditional approaches which evaluated trade-off between disclosure and utility, these approaches focus exclusively on privacy and provide a privacy guarantee regarding the ability of the adversary to identify an individual. Since an absolute guarantee of privacy is not feasible without making the data completely useless, these approaches allow for a small privacy leak. This leak or the privacy guarantee level (k in k-anonymity and ε in differential privacy) is chosen based on policy considerations. Once chosen, these guarantees ensure that the ability of the adversary will not exceed to identify an individual will not exceed the specified level. The resulting characteristics of the query response and/or released data is completely dictated by the privacy guarantee.

The attractiveness of the privacy first approaches is easy to understand; they provide a firm mathematical guarantee regarding the ability of the adversary to identify an individual. By contrast, traditional SDL assessments regarding privacy were typically based on an empirical evaluation of the ability of the adversary to identify an individual. In most cases, such an empirical evaluation required a model of the adversary. If the chosen model was that of a weak (strong) adversary, then an empirical evaluation of the chosen SDL technique using this model could result in high (low) privacy. The issue is further complicated since the SDL techniques explicitly consider the trade-off between disclosure and utility, sometimes adjusting the disclosure factor in favor of utility.

At the same time however, the definition of privacy in privacy-first approaches is very specific while SDL approaches take a much broader definition to include disclosure of identity, value, and even disclosure about vulnerable groups. Although these definitions may seem ad hoc (empirically based rather than theoretically based), they reflect a desire to assess disclosure from different perspectives. The primary objective of this paper is to raise the question: Does a privacy guarantee also translate to a guarantee against disclosure?