NATIONAL INSTITUTE OF STATISTICAL SCIENCES

SECURE STATISTICAL ANALYSIS OF DISTRIBUTED DATA

EXECUTIVE SUMMARY

This white paper is intended to lay out for the National Center for Education Statistics (NCES) available algorithms and technology solutions for secure, principled statistical analysis of distributed data, as well as to identify gaps in our current understanding.

For concreteness, the following problem served to frame the key issues. Consider a set of “similar” databases, such as statewide longitudinal data systems (SLDS), containing student-level data, in this instance, owned by multiple state education authorities (SEAs). For some reason – possibilities include law, regulation, policy, scale, distrust among the databases owners, lack of a trusted third party, and simple unwillingness – it is not possible to consolidate the data into a single database. Notwithstanding the restriction, the goal is to perform sound statistical analyses without having a consolidated database. Much of this white paper is devoted to explaining ways in which secure and principled analysis is possible, provided only that the database owners are willing to share anonymously certain summaries of their data, such as means and covariances.

The underlying literature spans both statistics and computer science, but the perspective here is statistical. First, the analyses must be *correct*, about which a computer science and statistics are in full agreement. Second, analyses must be *complete*, which is often not true in the computer science literature. Especially, those objects that statisticians use to quantify and characterize uncertainty must be provided. For a linear regression, for instance, in addition to estimated regression coefficients, the analysis must produce estimated standard errors, key test statistics and significance levels (“*p*-values”), information about model fit, model diagnostics and some information about residuals. Some of these objects can be produced using the same methods used to calculate estimators, but others cannot.

From a starting point of concepts and tools from computer science, the approach to secure statistical analyses is laid out using horizontally partitioned data. Brief discussion of privacy-preserving record linkage, analysis of vertically partitioned data, and complex partitions follows.

The paradigm described here and used to frame the problem suffers from a massive shortcoming in many practical applications: the database owners *must prescribe in advance what analysis is to be performed*. Exploratory data analyses are almost completely precluded, unless they are identified one-by-one in advance. Any adaptive human interaction with the data currently seems impossible.

Read the Full Report