The Cost of Provable Privacy: A Case Study on Linked Employer-Employee Data[*]

Samuel Haney[1], Ashwin Machanavajjhala[1], John M. Abowd[2], Matthew Graham[3], Mark Kutzbach[3], and Lars Vilhuber[2]

[1]Duke University, {`shaney,ashwin`}`@cs.duke.edu`
[2]U.S. Census Bureau, {`matthew.graham,mark.j.kutzbach`}`@census.gov`
[3]Cornell University, {`john.abowd,lars.vilhuber`}`@cornell.edu`

Abstract

In this work, we develop private mechanisms for releasing tabular summaries of employer-employee data. These summaries involve aggregating over the employment counts of sets of establishments. U.S. law mandates that the existence of an employee in the dataset and the precise employment size of an establishment be kept confidential. Direct applications of differential privacy that attempt to hide entire establishments result in poor utility. Instead, we identify the privacy requirements mandated by current interpretations of the relevant law, and develop a new formal privacy notion, related to differential privacy, customized to this specific application. Our notion guarantees that the employment of an establishment is protected to within a proportional factor α. Privacy of individuals is protected by standard ϵ-differential privacy. We design private algorithms and show that they have utility comparable to the existing ad-hoc protection system for an establishment-based data product published by the U.S. Census Bureau. A preliminary version of this talk was presented at the UNECE/Eurostat Statistical Data Confidentiality Work Session 2015.