[[1]](#footnote-1)

Methods for Anonymization of Large Data Sets for Public Release without Loss of Statistical Validity

Alex Deshowitz; Dennis Murray

*Abstract*—The broader release of large data sets from a multitude of sources has created new issues in privacy management. While organizations may be managing control of data internally better, the release of data sets for public scrutiny is opening new avenues to compromise of privacy and security. While the perceived risks of data release of seemingly innocuous data from taxi and Uber trips, or Netflix movie reviews, with personal data removed would seem to be safe, exposure to a wider audience with proficiency in data mining has exposed vulnerabilities. With public agencies releasing more data to attempt to harness advanced analytic techniques, ensuring privacy is paramount while still maintaining stastical validity. This research paper will explore the process for defining security requirements, as well as methods available via the Python programming language for reducing the traceability of personal data.

*Index Terms*—Anonymization, Privacy, Python

# INTRODUCTION

A

vailability of large data sets of previously privately held data has increased recently. The establishment of a Chief Data Scientist of the United States helped usher expanded availability of government data for public scrutiny. In private industry, the Kaggle website’s data mining competitions have established a method for private organizations to seek broader help with their data problems. The data sets published by public entities include crime data, as well as the results of the American Communities Survey (ACS) by the Census Bureau. Private data sets include information that could be considered highly sensitive, like anonymized health data. It also includes data sets that could be considered innocuous – like the data behind Netflix recommendation engine and the transactional records of New York City taxi cabs and Uber rides.

In both cases, the law of unintended consequences, and possibilities unforeseen by the owners of the data showed there was still personal data that could be harvested from the data sets.

The owners of the data have specific goals for release of their data, and thus must maintain adequate depth within the data for analysts to find statistically valid conclusions. However, they also want to allow for creative space in the data set – and the default is usually releasing a fuller, more complete data set over the release of a truncated or summarized data set. With interest in many data sets based in the variability, a fuller data set also serves to provide the depth required.

The topic has been documented previously by Latanya Sweeney in “k-Anonymity: A model for Protecting Privacy” (*International Journal on Uncertainty*, Volume 10 Number 5, 2002) in which the author models a system by which an individual would be undistinguishable from “at least k-1 individuals that also appear in the release.”

This paper will document the process of gathering requirements for maintaining private data, and then implement methods within an Python script to prepare a data file for release.

1. This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, “This work was supported in part by the U.S. Depart­ment of Com­merce under Grant BS123456”.

   The next few paragraphs should contain the authors’ current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@ boulder.nist.gov).

   S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

   T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp). [↑](#footnote-ref-1)