Institute of Transportation Studies
UNIVERSITY OF CALIFORNIA

Key Challenges in Sanitizing Transportation Data to Protect Sensitive Information

Matt Bishop, Ph.D.
Department of Computer Science, University of California, Davis

November 2021

Issue

As new mobility services such as ridehailing and shared micromobility have grown, so has the quantity of data available about how and where people travel. Transportation data provides government agencies and transportation companies with valuable information that can be used for identifying traffic patterns, predicting infrastructure needs, informing city planning, and other purposes. However, the data may also contain sensitive information that can identify individuals, the beginning and ending points of their trips, and other details that raise concerns about personal privacy. Even if a traveler's name and address is suppressed, adversaries could use other parts of the information such as trip origin and destination to learn an individual's identity and their habits. Similarly, another transportation company competing with the company that collected the data could potentially steal their customer base if they can use the data to obtain proprietary information such as frequent drop-off/pick-up locations, vehicle positioning, travel routes, or algorithms for assigning vehicles to clients.

To date, research has focused on how to *sanitize* records to protect sensitive information—by either omitting data or altering it—while still ensuring that analysis of the sanitized data produces the same results as analysis of the unsanitized raw data. However, many sanitizing strategies do not eliminate the threat of a *linkage attack*, in which the sanitized records are compared to data drawn from other public sources. If the data in the sanitized dataset can be correlated with data from another source, then it may be possible to reverse the sanitization. To better understand how to protect personal privacy and proprietary information while providing better data for transportation agencies, researchers at the University of California, Davis identified gaps in current sanitization strategies and questions that could lead to improvements in practice.

Key Research Findings

Even false or incorrect inferences drawn from sanitized data may be harmful. As an example, mileage and destination data showing an individual taking frequent trips to a hotel for business meetings could be falsely interpreted as evidence of an extramarital affair. There has been little research on determining and preventing incorrect inferences that could be drawn from anonymized data.

Sanitization must not affect the utility of the data. Analysis of sanitized data and the same analysis of the corresponding unsanitized data should produce identical, or at least "close enough," results to meet the purpose of the analysis (Figure 1). If the results are neither identical nor "close enough," then the sanitization is interfering with the utility of the data and must be weakened or the purpose of the analysis must be changed. Transportation agencies and companies must navigate this delicate balance between privacy and utility.

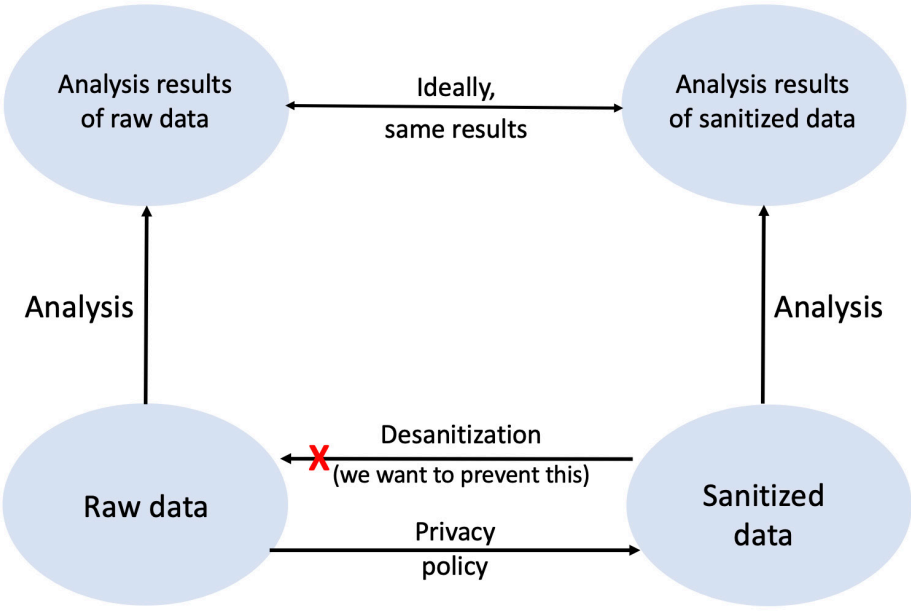

Figure 1. The sanitization process. A privacy policy dictates how raw data is sanitized to protect sensitive information (left lower oval to right lower oval). Both the raw and sanitized data are analyzed (arrows going from the lower ovals to the upper ones), and ideally the analyses should produce the same results (upper ovals and connecting arrow). The sanitization must be done to prevent desanitization (the top arrow between the lower ovals, and the "X" represents preventing this).

Preliminary work on linkage attacks shows that they can be effective in defeating data anonymization—the type of sanitization that removes personally identifying information from a dataset. Experiments in correlating anonymized transport data with overlapping social network data enabled most of the anonymized records to be deanonymized.[1]

Methods for determining what data will enable a linkage attack to work have not been developed. Some sources, such as social networks, contain data similar to transportation data. In this case, it is straightforward to determine what data could be used in a linkage attack and, potentially, take steps to prevent such an attack. In other cases, data not yet created or made publicly available may enable a linkage attack. How to determine what this data is presents a gap in current research.

More Information

This policy brief is drawn from the report "Sanitization of Transportation Data: Policy Implications and Gaps" prepared by Matt Bishop with the University of California, Davis. The report can be found here: https://www.ucits.org/research-project/2020-04/.

For more information about findings presented in this brief, please contact Matt Bishop at mabishop@ucdavis.edu.

[1] Srivatsa, M. and Hicks, M. Deanonymizing Mobility Traces: Using Social Network as a Side-Channel. In Proceedings of the 2012 ACM Conference on Computer and Communications Security, 2012. 628-637.

Research presented in this policy brief was made possible through funding received by the University of California Institute of Transportation Studies (UC ITS) from the State of California through the Public Transportation Account and the Road Repair and Accountability Act of 2017 (Senate Bill 1). The UC ITS is a network of faculty, research and administrative staff, and students dedicated to advancing the state of the art in transportation engineering, planning, and policy for the people of California. Established by the Legislature in 1947, the UC ITS has branches at UC Berkeley, UC Davis, UC Irvine, and UCLA.

Project ID UC-ITS-2020-04 | DOI: 10.7922/G2SJ1HXN