PRIVACY ANALYTICS
an IQVIA company

Summary Report: Re-identification Risks for Mobility Data

June 2020

This summary report was prepared for Uber. The summary results are specific to Uber's California Public Utilities Commission (CPUC) dataset and disclosure context. For more information on managing re-identification risk or for help developing safe data strategies, please contact **support@privacy-analytics.com** or visit **privacy-analytics.com**.

June 12, 2020

An Evaluation of Re-identification Risks for Uber’s California Public Utilities Commission (CPUC) Dataset

Executive Summary

In 2019, Uber contacted Privacy Analytics Inc. to assess the re-identification risk for a dataset, consisting of details on every ridesharing trip in California that Uber is required to submit annually to the California Public Utilities Commission (CPUC). We determined that the re-identification risk for Uber riders and drivers in the dataset was high, irrespective of contextual protections applied by data recipient or the data release type (non-public/public). We examined several reasonable de-identification strategies to reduce risk, including privacy measures such as data coarsening of times and locations, or suppression of rural or urban trips. None of the examined de-identification strategies or release types produced a meaningful reduction of the risk of re-identification. We conclude that with the given data transformations and disclosure contexts, the CPUC dataset has a high risk of re-identification of the individuals in the dataset.

This document provides a brief summary of the re-identification risk determination titled “An Evaluation of Re-identification Risks for Uber’s California Public Utilities Commission (CPUC) Dataset” (June 12, 2020). The complete report details the re-identification risk measurements, the considerations taken to determine an appropriate risk threshold, and the de-identification and masking steps that are recommended.

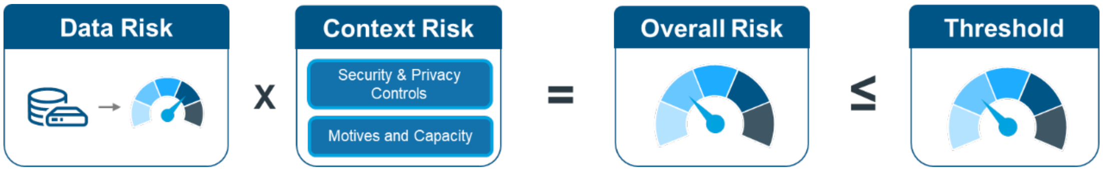
Background

For the CPUC Transportation Network Company Annual Data Report, Uber is asked to provide information for every Uber ridesharing trip in California. Since trip details are personal to the riders and drivers participating in these trips, Uber sought to understand to what extent this data could be used to determine an individual’s identity and personal trip information. Uber contracted Privacy Analytics Inc. to review the re-identification risk associated with the sharing of this data and provide guidance on how to protect future data releases from re-identification. Uber defined a data subset of the full data requested by the CPUC (here, the “CPUC dataset”), which included unique trip identification numbers, driver vehicle identification numbers (VIN), as well as dates, times and GPS locations for pickups and dropoffs. Extending the dataset to include additional fields is expected to increase (or, in the limiting case, have no impact on) the identifiability of the dataset assessed in this report. The independent re-identification risk assessment can be used to advise public officials and future collaborators on the inherent risks and protections necessary for this data.

Methodology Overview

We examined the risk of re-identification using the approach described below (Figure 1). Re-identification risk is modeled with two components: re-identification risk related to the distinctiveness of individuals in the dataset (the data risk) and re-identification risk mitigation related to the security and privacy controls safeguarding the data, as well as the motives and capacity of data recipients to attempt a re-identification (the context risk). The product of the data and context risks represents the overall risk of the dataset and

Figure 1. A schematic of the methodology approach



must be below an accepted threshold for the data to be considered to have a very small risk of re-identification.

The CPUC dataset is organized in terms of trips, but it implicitly describes details about Uber riders and drivers. Two fields (trip identifier numbers and vehicle identification numbers) were considered direct identifiers. These values are recommended to be masked, as they are keys to other databases and could easily be used to identify a rider or driver. Other fields (dates, times and GPS locations for the start and end of a trip) were considered quasi-identifiers: values that could be used individually or in combination to probabilistically re-identify a rider or driver. The data risk component was assessed for each individual as a measure of their distinguishability in the dataset, based on these quasi-identifying fields. These per-individual risk values were aggregated to define the data risk component of the risk measurement. Risk was measured on a two-month subset of the CPUC dataset.

The context risk component is a representation of the environment into which the dataset will be released. Practical attempts to re-identify data subjects are mitigated by the privacy and security controls applied by a data recipient and influenced by the motives and capacity of a recipient to attempt re-identification of the data. The CPUC was contacted to assess privacy and security controls but declined to share these details. Thus, this project considered all combinations of non-public and public data release contexts, including best-case assumptions.

The data risk and context risk components produce an overall re-identification risk for the dataset, and this is compared to a risk threshold. Determination of an acceptable risk threshold depends on several factors: 1) the level of invasion of the individuals' privacy based on the sensitivity and potential injury to the individuals in the dataset, and 2) past precedents of risk thresholds utilized across the industry [1]. For the CPUC dataset, an acceptable re-identification risk threshold of less than or equal to 0.057 was determined for a non-public release, while 0.056 was the determined threshold for a public release.

Uber requested that the CPUC dataset be measured in its original format of precise dates, times and GPS locations, as well as formats where Uber-specified de-identification strategies were implemented. These de-identification techniques include:

- **Coarsening the GPS field from a 5-decimal (precise) to a 3-decimal (coarse) GPS coordinate.** Note that Uber currently submits GPS coordinates to the CPUC with a three decimal point precision; therefore, this report examines the effect that this coarsening has on the re-identification risk of this dataset. Other jurisdictions where Uber operates do not permit truncated GPS values, so Uber requested an analysis at both three and five decimal point precisions to understand, more generally, the usefulness of this technique in the context of protecting privacy. This technique reduces accuracy by 100-fold, decreasing a value's distinguishability from other values.

- **Coarsening the time field from minutes (precise) to hours (coarse).** This technique reduces accuracy by 60-fold, decreasing a value's distinguishability from other values.
- **Releasing a subset of the dataset, such as only rural trips or only urban trips.** A dataset containing only pickup information was also examined.

The methodology used for risk measurement satisfies contemporary criteria for anonymization methodologies, and is consistent with the California Consumer Privacy Act, the U.S. Federal Trade Commission guidelines on data protection, U.S. Department of Health and Human Services (HHS) guidance, and other industry and global privacy standards, and has been publicly documented and peer-reviewed.

Key Findings

Non-Public Release: Re-identification risk is affected by the context of the data recipient. For example, a data recipient with strong privacy and security controls (e.g., well-defined data retention and destruction policies with periodic audits) and secure contractual agreements in place can protect against re-identification, while data recipients with minimal control and contractual agreements place provide only marginal mitigation of risks.

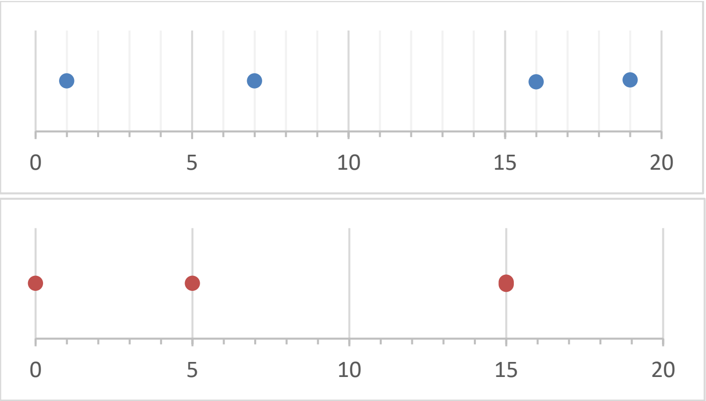
Table 1 summarizes the re-identification risk measurement results for the smallest context risk scenario (i.e., the scenario with the highest levels of recipient trust, as well as the highest levels of privacy and security measures in place). For this most permissive release context, a high risk of re-identification was measured for each de-identification strategy. **For every de-identification strategy and non-public release strategy examined, the CPUC dataset exceeds the permissible re-identification risk threshold of 0.057**.

Table 1. Summary of Re-identification Risk Determination on the CPUC Dataset for a Non-Public Release

De-ID Strategy	Privacy Techniques used in the Strategy	Dataset	GPS	Time	Pickup / Dropoff	Re-Id Risk
1	All data, full trip, precise location and precise time	All	5	Min	Both	0.14
2	All data, full trip, coarse location and coarse time	All	3	Hour	Both	0.14
3	All data, full trip, coarse location and precise time	All	3	Min	Both	0.14
4	Low Density, full trip, precise location and precise time	Low Density	5	Min	Both	0.14
5	All data, full trip, precise location and coarse time	All	5	Hour	Both	0.14
6	High Density, full trip, precise location and precise time	High Density	5	Min	Both	0.14
7	All data, pickups only, precise location, precise time	All	5	Min	Pickups	0.14

Impact of data transformations on re-identification risk: The methodology applied for this report defines risk as it relates to equivalence classes; that is, groups of individuals who are indistinguishable based on their quasi-identifiers. The dataset transformations that alter the accuracy (i.e., coarsening) or distribution (i.e., subsetting) of the quasi-identifiers should theoretically lower risk by making it more difficult to distinguish between simliar values. However, our analysis shows the CPUC dataset contains a substantial population of individuals that are modeled as being unique and remained unique despite the mitigating behavior generally expected of the data transformations applied. Figure 2 below illustrates this phenomenon for a single quasi-identifier.

Figure 2. An example of data coarsening



The top panel shows a quasi-identifier plotted to integer accuracy (*blue*), while the bottom panel shows the same quasi-identifier generalized to bins [0-4], [5-9], [10-14], [15-19] (*red*). We can see that even by coarsening the measurement from integers to five-integer bins that the first two points in this series remain unique.

If the sizes of the bins were to be increased by coarsening GPS or datetime to a greater extent than explored in this report, it is expected that at some level of coarsening the measured risk would decrease.

Public Release: In contrast to a non-public release, re-identification risk for a public release is not modulated by context protections. Futhermore, it can be assumed that an unintended party will inevitably attempt to access a publicly available dataset, and that this party will target the most distinctive individuals in the dataset, greatly increasing re-identification risk.

Table 2 summarizes the re-identification risk measurement results for the CPUC dataset for a public data release. **Every de-identification strategy considered for public release of the CPUC dataset is still over the permissible re-identification risk threshold of 0.056.**

Table 2. Summary of Re-identification Risk Determination on the CPUC Dataset for a Public Release

De-ID Strategy	Privacy Techniques used in the Strategy	Dataset	GPS	Time	Pickup / Dropoff	Re-Id Risk
1	All data, full trip, precise location and precise time	All	5	Min	Both	1.00

2	All data, full trip, coarse location and coarse time	All	3	Hour	Both	1.00
3	All data, full trip, coarse location and precise time	All	3	Min	Both	1.00
4	Low Density, full trip, precise location and precise time	Low Density	5	Min	Both	1.00
5	All data, full trip, precise location and coarse time	All	5	Hour	Both	1.00
6	High Density, full trip, precise location and precise time	High Density	5	Min	Both	1.00
7	All data, pickups only, precise location, precise time	All	5	Min	Pickup	1.00

The public release case, which lacks any contextual control, illustrates that individuals in the dataset are highly distinguishable on their indirect identifiers, even after substantial mitigating transformations. When additional mitigating contextual controls are applied to the model for non-public releases, the re-identification risk is reduced but still greatly exceeds the threshold, even when best-in-class release context is assumed.

Conclusion and Recommendations

The re-identification risk for the CPUC dataset was found to be high across every combination of de-identification strategy and release scenario considered. This finding pertains to conditions under which best-in-class context risk was assumed and direct identifiers were masked. This is primarily due to the substantial risk inherent in the data that can only be partially mitigated by strong privacy and security controls and coarsening techniques. In the absence of both - if the data was to be released to the public without dataset transformations - an even greater risk would be present. For these reasons, we expect that in every examined scenario, there is a high probability that the information contained in the dataset could be used by an anticipated recipient—alone or in combination with other reasonably available information—to identify an individual who is a subject of the information. This report demonstrates the need for more extensive privacy measures to transform the data in order to appropriately manage the risk of re-identification for the individuals described therein.

A potential alternative strategy for data release is to deliver aggregate-level data from the CPUC dataset. Several precedents for minimum acceptable size for an aggregate count (i.e., cell size) exist and often range from 11 to 20 individuals when data has detailed information, or publicly shared. For a single aggregate view, it is recommended that the minimum context-adjusted cell size should be 18. However, additional analyses should be conducted for overlapping aggregate deliverables.

Qualifications and Limitations

We confirm that:

1. The Risk Determination was conducted by qualified professionals with appropriate knowledge of and experience with generally accepted statistical and scientific principals and methods for rendering information not individually identifiable;
2. To the best of our knowledge, we have applied generally accepted statistical and scientific principles and methods for rendering information not individually identifiable in reaching our Determination; and
3. We have documented the methods and results of the analysis that justify our Risk Determination.

The statement set out above is subject to the following limitations:

1. The Risk Determination is based on the background information and sample data provided to us by Uber, and our Determination is contingent upon the assumption that such information is complete and accurate.
2. The Risk Determination is based on a documented series of assumptions which Uber has confirmed are reasonable to make given its business. These include (but are not limited to) the assumption that the CPUC dataset is representative of future Uber-generated datasets and that additional tables/fields would require a formal review by Privacy Analytics Inc. Additional assumptions are detailed in the full Re-identification Risk Determination report.
3. This Risk Determination is subject to all the terms and limitations set forth in the Engagement Letter between Uber and Privacy Analytics Inc., including all attachments that are incorporated therein by reference.

Our Risk Determination is valid for 18 months from the issue of the full report, provided there are no changes to the assumptions, the business of Uber, or the nature of the database (e.g., variables and distributions).

References

[1] K. El Emam, *Guide to the De-Identification of Personal Health Information*. CRC Press (Auerbach), 2013.

Authors:			**Reviewer:**
	Timal Kannangara Digitally signed by Timal Kannangara Date: 2020.09.11 12:12:30 -07'00'	Andrew Dickinson	Digitally signed by Jordan Collins DN: cn=Jordan Collins, o=Privacy Analytics, ou=Professional Services, email=jcollins@privacy-analytics.com, c=CA Date: 2020.09.11 15:21:39 -04'00'
Brian Rasquinha **Senior Solutions Architect**	**Timal Kannangara** **Data Scientist**	**Andrew Dickinson** **Data Scientist**	**Jordan Collins** **Data Science Lead**