www.csiro.au

CSIRO Submission 20/713

Government access to vehicle generated data

National Transport Commission

June 2020

Enquiries should be addressed to:

Ms Elizabeth Yuncken
CSIRO Government Relations
GPO Box 1700, Canberra ACT 2601
T 02 6218 3547
E mplo@csiro.au

Main Submission Author:

Dr Paul Tyler
Information Security and Privacy Group
CSIRO Data61

Introduction

CSIRO welcomes the opportunity to provide input to the National Transport Commission (NTC) on Government access to vehicle-generated data.

Technological changes are revolutionising road transport around the world, resulting in the potential to generate large amounts of data covering transportation and movement of people. There are many potential benefits from that data beyond its current use, especially if enabled through data sharing. There are also barriers to such data sharing. The NTC's Discussion Paper correctly highlights many of these areas.

CSIRO has world class research capabilities and delivers rigorous and comprehensive research in many areas that are directly connected to these technological changes including computer vision, robotics, wireless communication, cybersecurity and information security and privacy. This response draws from our experience in these areas and the data these systems produce, with a primary focus on privacy.

We provide some general comments on privacy, and then address selected questions from the discussion paper. CSIRO is available to discuss our response further if that would be of assistance to the NTC. Please refer to the contact details on the cover page.

General comments on privacy and data sensitivity

The NTC has discussed the issue of privacy and data sensitivity in the discussion paper. In section 5.2.4 (page 60), the discussion paper highlights the risk of reconstruction, or re-identification of individuals in location data. CSIRO has been involved in a few re-identification risk assessment tasks involving transportation data.

Generally, we have found that location and time information carry a significant risk of re-identification. Where the data is linked, multi-dimensional, or in the form of time series, not only is the risk magnified, but the ability to learn significant sensitive information about an individual is also increased.

Re-identification risk arises when information already known (background information) can be linked to data. Where the information already known matches only a small subset of the dataset, re-identification risk is significant. For example, a motorcycle with sidecar will be more likely to be re-identifiable in a dataset than a sedan.

Background information can also change the nature of data held from appearing non-sensitive and non-personal to being sensitive or personal. For example, if a location is held in the data for a particular event (crash, alert, CAM message) and someone knows that event occurred in front of the driver's home (but doesn't yet know where they actually live), then if they can re-identify the event and associate the driver with that event, then they can learn the address at which the driver lives. Background information can therefore turn what might seem like an innocuous location into something more sensitive or personal. Unfortunately, the data custodian cannot determine the risk that location could turn out to be sensitive and so might need to treat all such data as potentially sensitive or personal.

The risk of re-identification also tends to explode in longitudinal data or where it is easily linkable to other data. A set of time and locations tied by a common ID (even hashed) will be more re-identifiable than a single time and location. Such linked locations are very common in transport related data. Origin / destination pairs are very useful in planning and understanding population movement but carry privacy risk because of the possibility of being tied to individuals. For the primary purpose of collision avoidance, V2X CAM messages provide a constantly updating stream of location information, but carry privacy risk, especially when collected and stored in a central dataset. (This can hold even with many of the protections designed into such messages.) Also, data treatment methods may not remove all risk. Where an entity holds significant amounts of background information, aggregated data can reveal information about individuals through differences in the data over time, or between two datasets. Machine learning models can leak private or sensitive information of the underlying training data.

Ultimately privacy risk needs to be assessed in every data situation and then controlled either by data treatment or by security measures with known levels of residual risk. This task should not be underestimated in either the complexity or cost. CSIRO would encourage further understanding of privacy risk if government were to gain access or hold such data.

Response to selected questions in the Discussion Paper

Question 1: Do our problem and opportunity statements accurately define the key problems to be addressed, and do they capture the breadth of problems that would need to be addressed?

CSIRO encourages a broader consideration of the key problems be addressed. The problems listed in the discussion paper identify issues with access to data. However, access to data causes other risks to arise, particularly around privacy and security, and around data quality and trust. Some of these are touched on in the discussion paper. Here and in our response to Question 9, we present some additional issues that need to be considered.

In terms of privacy and security, unless careful analysis takes place to understand those risks, personal and sensitive information might be inadvertently revealed. Even in "de-identified data", a residual risk of re-identification may persist, and re-identification might be possible. A focus on the benefits of the data and of addressing the problems of access, may result in overlooking the risks that access might cause. In particular the issue of managing privacy and commercial confidentiality are potential major issues.

The issue of data quality and trust may also require further consideration beyond that presented in the discussion paper. Data quality is likely to be an issue in a system where data is sourced from different sensors or manufacturers. Where the data might inform government policy or decision making, there may be scope for a malicious actor to inject fake data or selectively filter data for their benefit. These risks should also be controlled.

Question 6: Is there value in establishing a national data aggregator or trust broker? Could good data definitions, practices and cooperation between entities achieve the same outcome?

As suggested in our response to Question 1, broader issues of privacy and security ought to be considered as part of assessing a solution involving a national data aggregator. A central data repository carries significant privacy and security risk. Those with access to the full datasets at the aggregator could potentially conduct linkage on the data sources to learn information about specific individuals. A single aggregator of data also makes that entity an attractive target for cyberattack.

Question 9: Have we accurately described the key barriers to accessing vehicle-generated data? Are there additional barriers?

In addition to the response to question 1, individual privacy and sensitivity, and commercial confidentiality can present barriers to vehicle-generated data, either due to the risks of holding such data, or due to individuals and other entities being unwilling to allow those risks to be taken up by other parties. For example, individuals may be reticent to allow access to data that might identify their place of residence. A vehicle manufacturer may be reticent to allow access to data that might reveal the locations where their customers reside or work. Those barriers can be difficult to break down, especially as techniques used in the past to manage these risks have not always been effective.

We present two case studies that illustrate unforeseen risks arising after access to data has been granted. In both cases, actual re-identification has been demonstrated based on publicly available background information. In both cases, data was "de-identified".

Case 1

In March 2014, the New York City Taxi & Limousine Commission (TLC) released data recorded by taxis' GPS systems in response to a freedom of information (FOI) request. The intended purpose was benign enough, to visualise the day in the life of a New York City taxi. The data consisted mostly of date, location and time of day for both the origin and destination of trips. The data became public as part of the FOI request. The data was "anonymised" and presumably considered "de-identified". However, just on the date, location and time of day data, a taxi is unique. (No two taxis can co-exist in the same location at the same time.) The data has since been used to illustrate how re-identification of taxis and their passengers can occur by linking the data with other known public information such as Facebook. At least two celebrities have been re-identified in the data. A detailed analysis has been conducted by Salinger Privacy[1], which also illustrates other privacy risks from this data.

Case 2

The myki travel card is used for public transport use in Victoria, Australia. Around July 2018, Public Transport Victoria (PTV) made a dataset of historical myki activity available for use in the "Melbourne Datathon". The data was "de-identified". Again, it wasn't long until a datathon participant raised concerns about being able to re-identify individuals in the dataset. Soon after, academics working at the University of Melbourne re-identify themselves and then a member of parliament[2].

The Office of the Victorian Information Commissioner (OVIC) investigated this release. CSIRO provided technical advice to the investigation. In the OVIC report[3], they noted the data contains information about individual's location at specific times, and assessed this to be "personal information". The de-identified data was found to be re-identifiable for a majority of the dataset, particularly on information of location and approximate time regarding two or more trips.

Both case studies illustrate that the risk of re-identification, especially in location and time-based data, can be overlooked or misunderstood, and this risk can be prevalent in transport related datasets. No information traditionally thought of as "personal" was required for these examples of re-identification. While the datasets were "de-identified" or "anonymised", these steps were not sufficient to prevent re-identification. They also illustrate the risk of ad-hoc de-identification methods where it is thought re-identification cannot occur but where there is no evidence or proof that the data is protected from re-identification.

We encourage NTC to reconsider the concept of "De-identified data" as being "information that cannot be re-identified", as this is rarely the case. Rather, we suggest that it be considered as information where the risk of re-identification has been reduced to a known and acceptable level.

The barrier illustrated here is one of unforeseen risk and the difficulty of controlling that risk. This is especially the risk of re-identification and impact on the privacy of individuals. Providing greater access to vehicle-generated data could potentially increase the likelihood of re-identification occurring on that data. Understanding and controlling that risk may be challenging and may impact the desired use cases of the data.

CSIRO would like to thank NTC for the opportunity to respond to their discussion paper. We would be happy to further discuss any of the issues raised in this response.

[1] https://www.salingerprivacy.com.au/2015/04/19/bradley-coopers-taxi-ride-a-lesson-in-privacy-risk/

[2] https://pursuit.unimelb.edu.au/articles/two-data-points-enough-to-spot-you-in-open-transport-records

[3] https://ovic.vic.gov.au/wp-content/uploads/2019/08/Report-of-investigation_disclosure-of-myki-travel-information.pdf