Paper number EU-TP0139

Anonymization of Data from Field Operational Tests

Yvonne Barnard[1*], Helena Gellerman[2], Sami Koskinen[3], Haibo Chen[1], Davide Brizzolara[4]

1. University of Leeds, Institute for Transport Studies, LS2 9JT, Leeds, UK

y.barnard@leeds.ac.uk; h.chen@its.leeds.ac.uk

2. SAFER Vehicle and Traffic Safety Centre at Chalmers University of Technology,

PO Box 8077, SE-402 78 Gothenburg, Sweden, helena.gellerman@chalmers.se

3. VTT Technical Research Centre of Finland, PO Box 1300, 33101 Tampere, Finland

sami.koskinen@vtt.fi

4. ERTICO – ITS Europe, Avenue Louise 326, B-1050 Brussels, Belgium UK, d.brizzolara@mail.ertico.com

Abstract

Data anonymization is a rising topic and is included in the EU's ITS directive. It brings new requirements for products and projects. Plans to arrange Field Operational Tests on connected ITS systems and automated driving are part of the European H2020 research and innovation programme. Data sharing will be an important part of those projects, aiming at widely studying the test data. Without anonymization and other data post-processing techniques to strip the logs of personal and confidential information, data sharing cannot happen or at least will require extensive non-disclosure agreements between several organisations. This paper discusses the state of the art on data sharing and anonymization techniques for FOT data sets, concentrating especially on video and GPS data.

Keywords:

Anonymization, Field Operational Tests, Personal Data

Introduction

Privacy protection in transport research is nothing new. Previously, data anonymization was primarily about removing names and addresses from single tables, and blurring faces and other identifying objects in still images. Today, with the internet search engines and various mobile phone applications collecting data, there is a need for stronger protection of privacy. Combination of data stored by several services can cause personal details to leak out. For example, one service continuously stores anonymous positioning data, and another service stores a single event with a location, time and also some identity information. The previously anonymous data about movement can now be linked to this identity data of the second application.

Mobile applications already collect a lot of data about their users. The users release some of their personal data, including position, in exchange for using services. Trust in proper data control is essential. The EU has revised the directive 95/46/EC from 1995, and a new regulation, mandatory for the member states, on the protection of privacy will soon be decided.

The automotive industry is continuously in transition. Currently, data-intensive services are prepared for the market. Cooperative Intelligent Transport Systems (C-ITS) services are based on data exchange between cars and roadside systems, where vehicle identity and precise position are exchanged for safety functions. Automated vehicles also involve continuous processing of video. The automotive industry is focusing on these new data issues, e.g., by appointing data protection specialists, and working together with legal experts and law-makers.

As an example of how personal data used to be anonymized in safety research we can look at the data about road accidents. In this data company names, positions, road signs, faces of people, including emergency personnel etc., are anonymized. However, where accident data uses still images, naturalistic driving study data includes continuous video. This, combined with the general trend towards open data and re-use of data, makes anonymization even more important. Legal requirements on privacy, participants who have only given consent to share anonymous data, and requirements in funding and consortium agreements, are some of the driving forces for new techniques for anonymization.

This paper discusses the state of the art on data sharing and anonymization techniques for Field Operational Tests (FOT) datasets, concentrating especially on video and GPS data.

Personal data in Field Operational Tests

In FOTs large amounts of data are gathered in order to study driving behaviour in interaction with ITS. In FOTs vehicles are instrumented with sensors to register what is going on inside the vehicle and in the traffic environment. During the last decade hundreds of these studies have been performed, involving thousands of drivers. In the near future large-scale FOTs will be performed with C-ITS and automated vehicles.

Carrying out a FOT usually means asking participants to give insight into their safety and mobility behaviour. Although participants may be given some form of (financial) compensation, or in the case of fleet drivers it may form part of their job, usually participants join in order to help advance road-safety, mobility and innovation. This means that the protection of the privacy of participants is extremely important, guaranteeing that their personal data is protected against use for which the participants did not give their consent. There are legal rules, at national and European levels, addressing data protection, but equally important are the underlying ethical principles, defined as respect for the person and his or her autonomy, dignity and self-determination. Human rights

legislation is also relevant, as is the Helsinki Declaration of 1964 and its subsequent revisions. This declaration enshrines the right of the individual to be informed and provide prior consent on a voluntary basis. Next to care for participants, the rights of non-participants, affected by the FOT, should be taken into account. These may, for example, be passengers in the participating vehicle or other road-users who are videoed.

Three categories of sensitive data can be distinguished: personal, commercial, and research sensitive data that may compromise the goals and objectives of a research endeavour.

In summary, there are several reasons for data protection and anonymization:

- Requirements from legal authorities and ethical committees, in order to get permission to conduct the study;
- Ethical reasons, to protect participants;
- Requirements from funders;
- Commercial reasons, as data may have a value for companies;
- Ensuring data is not used abusively, endangering current and future studies.

Data protection during and after the FOT

Protection of data of participants starts with the participant agreement, signed by the participant, and describing the purpose of the study and the use that will be made of the data, usually defined as use for (some form of) transport research.

When the FOT starts, the identity of the participant and his/her details are known by the organisation that is responsible for conducting the FOT. The data that is collected is assigned a driver ID, and the link between identity details and the data collected is only known by named persons, and is stored securely.

During data collection, location data is usually gathered and is frequently also combined with video, both allowing identification of the participant. Data should be stored using sufficient protection for the personal data, and only authorised personnel should be able to link the collected identity details to the participant, for a limited period of time.

During the data analysis analysts will work with the participant ID, not with names. As long as data remains within the consortium responsible for the FOT, protection of personal data can be well-organised. When data is shared with other organisations, and by other researchers to answer new research questions, this may become more difficult. Sharing is very important to enable optimal use of data usually collected at great cost, and to gain as much knowledge as possible about, e.g., crash causation and the potential impact of large-scale introduction of ITS and automated vehicles. If data remains in a secure environment, the whole dataset could be shared. However, sharing data publicly can only be done if datasets are fully anonymous, and abuse of personal data is ruled out.

Main issues in data anonymization

In general there are well-established procedures for anonymizing data such as the personal details of participants. However, in naturalistic data there are two major areas where this is problematic. The first is video data, and the second vehicle location data.
Video data may provide information about:

- Driver;
- Passengers;
- Other road-users, such as pedestrians, cyclists, drivers of other vehicles;
- Number plates of other vehicles;
- The environment, providing information about the location of the vehicle.

An example is the Naturalistic Driving Study UDRIVE, in which seven cameras gather video images both from the inside the vehicle and the surroundings (UDRIVE; Eenink et al, 2014)

Video may provide location data, but GPS sensors and map data, in combination with data about trip length, stops etc. can also reveal where the driver is going. These data may be sensitive because:

- Patterns in mobility behaviour may provide information about the location of home, work, school, etc. of participants. If, for example, a vehicle starts and stops each workday at the same place, we can assume that it is the daily commute. These patterns may reveal a lot about the daily lives of people.
- The driver may be located at a place which he/she would rather not reveal to others.

It will be clear that anonymization of these data types is a very desirable goal. However, next to technical difficulties to achieving it, another main question is how to preserve the original richness of the data: anonymizing data without losing essential information. Related to this is the question of when the anonymization should take place, at the source as anonymization by design, during data collection in the vehicle, when stored, as a pre-phase for analysis or before data is shared. One potential problem to take into account is that if the personal data is anonymized early in the process, before the original videos can be stored, the knowledge about which essential information to preserve becomes crucial as the process is irreversible.

Anonymization of video data

There are several techniques for anonymizing video images of the driver. Common image anonymization options are blurring, pixelation, bar mask, and negative colouring.

When considering the driver's face in automotive research, blurring destroys much information that could be valuable. For example, when analysing an event where a collision was about to happen, the

expression on the face of the driver provides a lot of information. For example, did the driver look anxious, being aware of a dangerous situation or was he/she relaxed.

Where these techniques obscure the driver's face, newer techniques try to preserve the essential information while anonymizing the characteristics of the face that allow identification but are not needed for the study's purposes, such as the form of the nose and mouth. An example of de-identification in the automobile environment, which uses such a new approach, is trying to retain as much information as possible by preserving head pose, gaze and facial expression, while removing person-specific facial features through transferring them to the face of someone else. This "someone else" might be a real person, like one of the researchers, a face composed of a combination of several faces or a synthetic person.

A different approach is to identify different facial movements (by muscles around eyes, mouth, brows etc.) and code them as separate units, e.g. upper lip raised being one "action unit". Facial actions are related to expressions, showing different types of emotions. There is a standard Facial Action Coding System (FACS) which has been in existence for a long time and is often updated. There are databases of these Facial Actions Units and automatic detection software. FACS is used in psychological research and in the animation industry (Ekman & Rosenberg, 1997).

When expressions are coded automatically they can be used to project the same expressions on other faces (human or avatars) or the researchers may use the codes and the expressions they represent directly, without the need to see video images.

The Federal Highway Administration (FHWA) runs a programme on video analytics research, with six projects developing different approaches for data processing and analysis, working on the automation of video data decoding (FHWA, 2016).. These techniques are targeted to make the huge SHRP2 dataset (the US naturalistic driving study) more easily available for analysis (SHRP2). Two of these projects are DMask and DCode (Tamrakar, 2015), providing anonymization by coding driver activity and driving context (DCode), and by masking the driver's face and body (DMask). The approach to anonymization consists of three tiers: core feature extraction by tracking relevant features such as faces and upper body position, intermediate feature extraction by monitoring and analysing gaze, expression, gestures etc., and feature integration, leading to the final coded features of the driver actions, state and environment. The masking approach masks out the driver's head with an overlaid synthetic avatar. Both approaches make use of learning mechanisms to improve performance. Although results are good so far, the success of facial motion transfer and identity masking is totally dependent on the accuracy of the facial feature tracking. And to have 100% accuracy, a human analyst still has to be involved in the process.

So far much emphasis is put on the anonymization of the driver. However, anonymization of the images outside the vehicle is also becoming more of an issue. As other road users never gave their consent, their privacy should be even more strictly protected. Additionally, even when all people and vehicles on video are anonymized, the environment is not. By looking at the images of the surroundings of the video, where a participant is driving may be easily identifiable, for example, by looking at easily identifiable landmarks such as monuments and shops.

The need for anonymization is depending on the quality of the video image, to what extent details can be recognized. The anonymization technique used for other road-users, number plates of other vehicles and easily identifiable landmarks is usually blurring. However, quality depends on the techniques for recognising them.

Anonymization of vehicle position data

While identification of the identity of people on video may be an important privacy issue, the location of their vehicle and their mobility patterns may pose even bigger challenges, as from the locations people may again be identified, and the anonymization of GPS traces is difficult. Tracing people's position by GPS is no problem, and part of the normal data collection from vehicles and smart phones. A large portion of all data generated today is transmitted through mobile networks. They capture locations, but users are often unaware of the collection and use of this information. Location traces could be a goldmine – but privacy is a huge concern. Simple anonymization seems not to be enough either. De Montjoye et al (2013) showed that four observations from cell phone data are sufficient to identify 95% of individuals.

Identifying mobility patterns poses a dilemma: they are extremely important for studying the impact of all kinds of traffic measures and intelligent transport systems, e.g., on congestion and road use, but they provide a wealth of data identifying persons. A normal everyday pattern, such as starting from home, driving a child to school, a short stop, driving to work, parking for 8 hours, driving to a supermarket, parking for 20 minutes, and driving home again is easily recognisable from GPS, map and drive/stop data. This shows exactly where participants live and work, whether they have children and which school these attend, when something out of the ordinary happened (holidays, illness) etc. etc. From a few simple data collected in a FOT, insight into the private life of participants can be deducted, as well as their identity that might have been so carefully anonymized in the video anonymization.

Traditional methods to anonymize trajectory data include masking true trajectories, suppressing trip origins and destinations, and redefining the geospatial reference layer by offsetting latitude/longitude pairs (e.g. all trips start from 0, 0). These techniques are effective in masking Personally Identifiable Information (PII) but limit data utilization.

One of the techniques is truncating start and end part of trips. In the Research Data Exchange (RDE), the US DoT transportation data sharing system, an iterative process is used (Henclewood, 2015). Intervals of concern are identified (e.g. stops), extended to privacy intervals, and the intervals are finally eliminated. Detection is done by using different features and methods, for example, the use of reverse and park can often indicate the stop/end location as well. GPS quality has been an issue, as poor positioning can cause problems for determining the value of distance and time parameters.

An important concept is k-anonymity, requiring at least k records for every possible value of any subset of attributes (e.g. k persons for any gender and date of birth). Anonymization can also be done by generalising a value in order to make it less specific, e.g. age 34 becomes 30–40; suppression by simply deleting the value; and perturbation, replacing the actual value with a random value out of the standard distribution of values for that attribute. Also l-diversity is required: diversity of the sensitive attribute within the k-anonymity set (Gidofalvi, 2015).

One indirect method of vehicle trace anonymization is making partial selections from the original dataset, covering merely the periods relevant to a study. For example, analysts see drivers' behaviour at specified locations, not their full trips. Generally, post-processing a GPS dataset offers further steps for hiding the original data: it is possible to generate summary datasets of driving, listing e.g. drivers' average speed and total distance driven on a certain speed limit area. When such summary calculation processes are first applied to data from hundreds of drivers, the next phases of analyses become statistical work – instead of studying actual individuals. When individual behaviour of interest, e.g. slowing down after receiving a warning, is summarized using a numeric indicator, an image of the driver is not needed.

Topics to consider in anonymization

It is very difficult to anonymize all data 100%; even if one type of data is completely anonymous, combination of data (car type, colour, position etc.) can point to one individual. Even if legal requirements are met, what people consider as private and do not want to be disclosed is very personal. An option in FOTs is to make an anonymization on an individual level, through asking persons what they do not want to have revealed, for instance, which GPS positions they do not want to reveal. Trying to make data as secure as possible, and to communicate about how data is protected and used according to good scientific practices, and the benefits of studies, is probably the best way forward.

The issue anonymization has to avoid is re-identification. Even if humans can no longer re-identify a participant, a computer may. Computers are able to recognize faces by using various measurement and search techniques. Background search may also be an issue, e.g. searching the internet for photos that

have the same background at the same date. The driver may have posted a photo of him- or her- self online, with that background.

To answer the question 'when is anonymization good enough?' we need to know how much anonymization is really needed; research, ethical and legal requirements are not always aligned.

There is not yet enough experience with how well researchers can do their analysis with anonymized data. Combination of techniques may also make analysis easier. We need to know more about what researchers require, how accurate the marked data should be, and what error rate and false alarms rates are acceptable. How much information is needed to understand the driver and the situation? For example, when using avatars, is gender and ethnicity important? There is a lot of emphasis on extracting facial features, but other features, such as body language are also important to understand what is going on. In data-collections such as SHRP2, cameras do not always capture other body parts very well.

Users of the anonymized data, the analysts, must be informed about the quality of processing, e.g. what expressions carry over accurately, and which expressions may become mixed up. It is important to understand the reasons for anonymization algorithms' failures, e.g. low quality of the original video. The models that are used should be adaptable. New sensors are going to help to improve data quality, such as 3D visual data sensing and infrared sensors for bad lightning conditions.

There is a balance between saving and deleting data for privacy reasons. Communication and transparency about what can be made public and what not, and the reasons for this, are crucial.

Conclusions

New and promising techniques are emerging but not all problems related to anonymization have been solved. The anonymization of driver faces is rather advanced, but context is more difficult, and anonymizing location and trajectories is still problematic. The latest anonymization techniques still have to prove their value for the large FOT datasets. If we want to use them for automated driving products and (car-to-car) communication systems instead of new research, perfection is needed. Original data (and metadata) quality is a key for being able to apply the techniques successfully.

Anonymization is becoming more and more important, due to a growing awareness of the need to protect participants' privacy, new laws and regulations, and requirements from project funding bodies. International collaboration and discussion between researchers, computer scientists and legal experts is necessary to advance the development towards more and more successful ways to make personal data anonymous. If the data could be anonymized while preserving the information that is essential for research, the access to and re-use of valuable data would be greatly facilitated.

Acknowledgement

Part of this paper is based on presentations and discussions at the FOT-Net Data workshop on Anonymisation of personal FOT data, 01-02 September 2015 in Gothenburg. FOT-Net Data is a Support Action in the seventh Framework Programme Information and Communication Technologies. It is funded by the European Commission (EC), DG Connect, under Grant Agreement number 610453.

References

Eenink,R., Barnard, Y., Baumann, M., Augros, X., and Utesch,F. (2014) UDRIVE: the European naturalistic driving study. Proceedings of the Transport Research Arena, Paris.

Ekman, P., & Rosenberg, E. L. (1997). What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA.

FOT-Net Data workshop on Anonymisation of personal FOT data (2015) [Online]. Report and presentation available at: http://fot-net.eu/Documents/workshop-on-anonymisation-of-personal-fot-data-01-02-september-2015-in-gothenburg/ (Accessed 18 April 2016).

FHWA (Federal Highway Administration) (2015). Exploratory Advanced Research Program: Video Analytics Research. Available at: https://www.fhwa.dot.gov/advancedresearch/pubs/15025/15025.pdf (Accessed 18 April 2016).

Gidofalvi, G. (2015). Trajectory Privacy: Measures and Preservation Methods. Presentation at the FOT-Net Data workshop on Data Anonymisation, Gothenburg. Available at: http://fot-net.eu/Documents/workshop-on-anonymisation-of-personal-fot-data-01-02-september-2015-in-gothenburg/ (Accessed 18 April 2016).

Henclewood, D. (2015). Anonymization of Traffic Data: Balancing Utility and Privacy - the U.S. Perspective. Presentation at the FOT-Net Data workshop on Data Anonymisation, Gothenburg. Available at: http://fot-net.eu/Documents/workshop-on-anonymisation-of-personal-fot-data-01-02-september-2015-in-gothenburg/ (Accessed 18 April 2016).

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3.

Research Data Exchange (RDE) website (2016) [Online]. Available at: https://www.its-rde.net/ (Accessed 18 April 2016).

Strategic Highway Research Program 2 SHRP2 [Online]. Available at: http://www.fhwa.dot.gov/goshrp2/Solutions/All/NDS/Naturalistic_Driving_Study (Accessed 18 April 2016).

Tamrakar, A. (2015). Coding Driver Activity (DCode) and Masking Driver's Face/Body (DMask) techniques. Presentation at the FOT-Net Data workshop on Data Anonymisation, Gothenburg. Available at: http://fot-net.eu/Documents/workshop-on-anonymisation-of-personal-fot-data-01-02-september-2015-in-gothenburg/ (Accessed 18 April 2016).

UDRIVE (European Naturalistic Driving Study) official website [Online]. Available at: http://www.udrive.eu/ (Accessed 18 April 2016).