Deliverable 2:
Anonymisation case studies

ODI Tender RDP8-001

Authors: Gemma Galdon Clavell and Carmela Troncoso

Research assistants: Victoria Peuvrelle and Miguel Vallbuena

Eticas Research and Consulting

C/ Ferlandina, 49

08001, Barcelona

(+34) 93 600 54 00

www.eticasconsulting.com

Table of contents

Health data

The General Data Protection Regulation classifies certain data as *sensitive* and deserving of specific precautions and regulations (see article 9.1 General Data Protection Regulation - GDPR[1] from now on). Sensitive personal data includes data that reveals racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for uniquely identifying a natural person and data concerning *health* or a natural person's sex life and/or sexual orientation.

Therefore, health data is a special category of data deserving specific precautions and safeguards regarding re-identification risks. Anonymisation failures in health datasets raise special concerns, as the medical history of an individual may have implications on their ability to secure a job, for instance, on their insurance premiums or on their life chances in general. Moreover, health data also includes biometric identifiers that, unlike name or an address, a person can't just change. Data breaches in this area, therefore, may cause permanent damage.

The most well-known case of re-identification of individuals through supposedly anonymised data in the field of health is the one where Sweeney (2000) managed to identify the hospital visits of Bill Weld, then governor of Massachusetts. She correlated the anonymised records of state employees' hospital visits with the publicly available voter roll of the city where the governor lived. By matching the ZIP code, sex and birth date, she managed to re-identify his record (Ohm, 2009). The entity in charge of releasing the data, the Group Insurance Commission of the Government of Massachusetts, had removed direct identifiers from the dataset, but left in the above information, which acted as quasi-identifiers and, when combined, led to re-identification. Their error was assuming that only direct identifiers could identify individuals. Sweeney's work thus showed the weakness of anonymisation in the case of linkage attacks.

A similar re-identification case occurred in 2017 in Australia, where the Australian Medicare Benefits Scheme (MBS) and the Pharmaceutical Benefits Scheme (PBS) released patient records that had, in theory, been anonymised.[2] As in the case of Sweeney, the methods used were not robust enough and researchers from the University of Melbourne found that the individuals could be re-identified by matching unencrypted parts of the datasets with previously known information about specific individuals, such as their gender or particular medical interventions that had been performed on a particular individual.

---

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679 [Accessed 22 Aug. 2018].

[2] Grubb, B. (2018). Health record details exposed as 'de-identification' of data fails. The Sydney Morning Herald. [online] Retrieved from: https://www.smh.com.au/technology/australians-health-records-unwittingly-exposed-20171218-p4yxt2.html [Accessed 22 Aug. 2018].

These cases point to the need to improve the way health data is sanitised to avoid re-identification when publicly releasing data. However, it is important to highlight that anonymisation is just one step in the chain of responsible data decisions that need to be taken when dealing with personal data. Issues of security and governance can also arise in the health field. For instance, the Information Commissioner's Office (ICO), the UK data protection authority, found in 2017 that the personal health records of 1,6 million patients of UK's National Health System (NHS) had been illegally handed over to DeepMind, Google's artificial intelligence firm. The information was shared to help create an app aimed to alert, diagnose and detect acute kidney injury. According to the ICO, the NHS had failed to comply with data protection law because patients were not aware of how their data was being used.[3] In this instance, transparency, consent and proper notice are also important issues.

Another instance where issues different than anonymisation were at the root of privacy failures also took place also within the NHS. In this case, data from patients that had requested for their personal health information to only be used for primary purposes (medical care) was shared with third parties for secondary use due to technical problems posed by the software used by doctors to register the patient's choice in this respect.[4] A new opt-out system, which does not require the intervention of the doctor, was launched in 2018.[5] In 2013, the NHS had also launched an opt-out option in the context of the creation of an online platform centralising patients' medical records. The programme was shut down amidst controversy surrounding the NHS not respecting patients' opt-out choices.[6] All of these examples illustrate how matters related to governance and data management can be as harmful to the privacy of data subjects as poor anonymisation procedures.

A specific case in point is that of genomic data. Using genomic data in research is an area of concern due to the possible negative consequences of losing control over such personal information, heightened by the stability in time of that data (Wang *et al*., 2009; Homer *et al*., 2008). Homer *et al.* (2008) developed a theoretical framework for detecting an individual's presence in a complex genomic data mixture, demonstrating it was possible to identify a participant in a genomic study because of the participant's allele frequency in large numbers of populations' DNA frequency variations. Wang *et al.* (2009) also conducted two attacks and

---

[3] Royal Free - Google DeepMind trial failed to comply with data protection law. (2017). Retrieved from https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/ [Accessed 22 Aug. 2018].

[4] BBC News. (2018). NHS data breach affects 150,000 patients in England. Retrieved from https://www.bbc.co.uk/news/technology-44682369 [Accessed 28 Aug. 2018].

[5] National data opt-out programme - NHS Digital. Retrieved from https://digital.nhs.uk/services/national-data-opt-out-programme [Accessed 22 Aug. 2018].

[6] Ramesh, R. (2015). NHS disregards patient requests to opt out of sharing medical records. *The Guardian*. Retrieved from https://www.theguardian.com/society/2015/jan/22/nhs-disregards-patients-requests-sharing-medical-records [Accessed 28 Aug. 2018].
Heather, B. (2016). Data breaches ongoing as NHS Digital pushes opt-outs. *Digital Health*. Retrieved from: https://www.digitalhealth.net/2016/09/data-breaches-ongoing-as-nhs-digital-pushes-opt-outs/ [Accessed 28 Aug. 2018].

were able to identify individuals from even smaller sets of data. Humbert *et al.* (2013) and Shringarpure Bustamante (2015) respectively created two tests to assess the likelihood of re-identification.

- **Health data in Data Protection legislation**

Personal health data is considered *very sensitive* in GDPR, and as such it must be treated with special care. The General Data Protection Regulation in the EU (GDPR) defines in article 4 "data concerning health" as data "related to the physical or mental health of a natural person" and links genetic data to health when such data "give[s] unique information about the physiology or the health of that natural person".

Both health and genetic data are considered as "special categories of personal data", akin to data revealing racial traits, political opinions or sexual orientation. These types of data "merit higher protection" (recital 53) and their processing is thus *prohibited* under article 9. Paragraph 4 gives member states the right to further legislate on the matter of genetic data, biometric data or data concerning health, taking the GDPR as a minimum standard.

There are exceptions allowing the processing of these "special categories of personal data", listed in article 9(2) and article 89, including:

- When consent was given by the individual for specific purposes
- In the context of employment, social security and social protection law
- To protect the vital interests of the individual
- When the data has been made public by the individual
- When the data is required by a court of law
- For reasons of public interest (national or international)
- For scientific or historical research purposes
- Archiving purposes in the public interest
- Statistical purposes

The case of health data is also specifically addressed in recital 53, where the summary of the exceptions is described as such:

> *Special categories of personal data which merit higher protection should be processed for health-related purposes only where necessary to achieve those purposes for the benefit of natural persons and society as a whole.*

In such cases, the security of the data must be specially safeguarded, taking into account multiple contingent factors detailed in article 25, including the following:

- The state of the art
- The cost of implementation of the measures adopted
- The scope, context and purposes of processing
- The risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing

Another relevant piece of EU legislation related to health data is Directive 2011/24/EU[7] on the application of patients' rights in cross-border healthcare. As recital 25 argues:

> *Ensuring continuity of cross-border healthcare depends on the transfer of personal data concerning patients' health. These personal data should be able to flow from one Member State to another, but at the same time the fundamental rights of the individuals should be safeguarded.*

To this effect, the Directive makes member states responsible for the processing of personal data to be compatible with the fundamental right to privacy, in conformity with GDPR.

A different framework worth mentioning is that of the European Medicine Agency, the scientific agency in charge of the evaluation of medical products. Effective since 2015, the agency published a policy on the publication of clinical data for medicinal products for human use (Policy 0070). [8] The policy addresses the issue of data protection by stating that any data processing must be fully compliant with Directive 95/46/EC[9] (and now with GDPR). Interestingly, the policy also raises concerns about the increasing difficulty to anonymize, and about potential future re-identification risks of anonymised data due to "data mining and database linkage". Consequently, the Agency "takes a guarded approach to the sharing of patient-level data".[10] To balance privacy and scientific utility, the agency has taken into account technical developments, and opinions and needs of stakeholders to decide on the optimal approach to anonymisation versus pseudonymisation. Its external guidance, published in 2017, dedicates a full chapter to anonymisation of clinical reports and is a key instrument in this field. The guidance reminds entities that no personal data of trial participants should be published. The same applies to personal data of clinical studies personnel, with the exception of the main investigators. The guidance provides recommendations on how best to achieve anonymisation

[7] Directive 2011/24/EU of the European Parliament and of the Council of 9 March 2011 on the application of patients' rights in cross-border healthcare. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32011L0024 [Accessed 22 Aug. 2018].
[8] European Medicine Agency (2017). *External guidance on the anonymisation of clinical reports for the purpose of publication in accordance with EMA Policy 0070*.
[9] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31995L0046 [Accessed 22 Aug. 2018].
[10] European Medicine Agency (2014). *European Medicines Agency policy on publication of clinical data for medicinal products for human use*.

and covers the procedural aspects and requirements related to the submission and publication of clinical reports.

- **Protecting health data**

The principal goals of health data collection and analysis are twofold: providing a better and more personalised assistance to the patient (primary use) and allowing for research to be carried out (secondary use) (Fernández-Alemán *et al*., 2013). In health data, it is clear that anonymisation often cannot happen at the time of data collection, as the primary use relies on personal data. Data minimisation also has limits, as some health issues are related to contextual and life choices (sexual habits, smoking, etc.) that need to be recorded to improve diagnosis. It is also important to mention that anonymisation may hinder health research or go against fundamental ethical principles in the health field, limiting the choice of anonymisation techniques at hand (Iacono, 2007). Finally, the permanent nature of some of the personal data collected (biometric traits, genomic data) also contributes to making this field the most complex and challenging when it comes to anonymising data to release it for further use.

Therefore, while it is important to have robust data quality and data minimisation measures in place, when sharing personal health data *inside* an organisation the key remedial approach will be creating an accountable and secure governance, sharing and access framework that ensures that the relevant data gets to whoever needs it including all the necessary information for a given context.

- A secure governance, sharing and access framework

Health institutions manage large, sensitive datasets for different purposes. A doctor and a clerk may work at the same hospital, but performing their duties requires different levels of access to data. Information on the medical history of a patient may not be necessary for accounting purposes, for instance, and while all the data may be in the same location, access control can ensure that one can only access the minimum information needed to perform their role in the organisation. The most common access control system used in hospitals and health institutions is ''Role-based Access Control'' (RBAC) (Fernandez-Aleman *et al.* 2013, p. 552). RBAC covers the security needs of organisations by associating permissions to administratively assigned roles in an organisation (Ferraiolo, Cugini, & Kuhn, 1995). It has been argued that this governance model allows for a sufficient degree of flexibility within an organisation in terms of the distribution of roles and, at the same time, provides high standards of security (Ferraiolo & Kuhn, 1992).

The ''My Health Record'' system in Australia, a database that contains personal health information from Australians,[11] is a good reference in this respect. It combines elements of the RBAC model with features that enable users to make decisions regarding how their data will

[11] What is My Health Record?. Retrieved from https://www.myhealthrecord.gov.au/for-you-your-family/what-is-my-health-record [Accessed 22 Aug. 2018].

be accessed and by whom, thus incorporating a layer of user control over the data stored. In this system, even though in principle every registered healthcare provider can access a patient's profile, there is an option to generate a password that will be required to access the files and that only the patient will know.[12] Even though the system was initially controversial due to concerns related to the user's ability to opt out of the system and to the management of sensitive personal information,[13] a report from the Office of the Australian Information Commissioner in 2018 showed the system is not as vulnerable to data breaches as some claimed. It is however important to note that the default setting does not include the password-protected option.[14]

For sharing data for research purposes, Ohm (2009) suggests the creation of legal protocols on data sharing and processing that are not so technology-based but more reliant upon governance frameworks and trust. The codification of such rules could however prove difficult and should be backed by further safeguards and accountability measures. Particularly interesting is the idea that access to data could be regulated according to its level of sensitivity. The most sensitive data (for instance, HIV testing results, genomic data or mental health records) could have a specific regime (for example, making the researcher physically go where data are being stored), different from other datasets with less sensitive information (which could be accessed remotely, generating logs to identify who accessed what and when). The rationale behind proposals such as this is that adequate governance and access frameworks could better protect privacy rights while, at the same time, enable better research as there would be no need to perform anonymisation techniques that would inevitably compromise the quality of the datasets. In any case, it is worth noting that today most of the emphasis in protecting health data is not on governance and access frameworks but on technical solutions (Ohm, 2009, p. 1769).

Finally, in the case of genomic data, third-party encryption systems stand out as a feasible and effective option to protect genetic data while enabling re-identification in cases where research or ethical reasons make it advisable to do so (Gulcher *et al*., 2000). However, this is still a novel approach, and current encryption systems tend to not involve third parties whatsoever. Instead, the providers keep the encryption keys in their own servers. Some other approaches advocate for patients being able to generate their own encryption keys (as in the Australian ''My Health Record'') in order to mitigate re-identification risks even more by deciding who will be able to

[12] My Health Record: Frequently asked questions. (2018). Retrieved from https://www.myhealthrecord.gov.au/for-you-your-family/howtos/frequently-asked-questions [Accessed 22 Aug. 2018].
[13] Grubb, B. (2018). Australians are 'rightly' questioning My Health Record, says Privacy Commissioner. *The Sydney Morning Herald*. Retrieved from https://www.smh.com.au/technology/australians-are-rightly-questioning-my-health-record-says-privacy-commissioner-20180730-p4zui3.html [Accessed 28 Aug. 2018].
[14] Howell, B. (2018). Data privacy debacle down under: Is Australia's My Health Record doomed? [Blog]. Retrieved from http://www.aei.org/publication/data-privacy-debacle-down-under-is-australias-my-health-record-doomed/ [Accessed 22 Aug. 2018].

re-establish the link between the data points, although the success of this procedure hinges on the assumption that the patient will store the code safely. In order to mitigate the risk of third parties getting hold of the codes, encryption systems based on more layers of security such as wireless mobile devices have been proposed and implemented (Sax, Kohane, & Mandl, 2005).

The need for secure, ethical and responsible governance, sharing and access frameworks is relevant in all cases, regardless of the kind of data to be used and the further uses foreseen. Indeed, many data protection issues can be avoided without complex technical solutions or techniques -robust consent and security mechanisms may suffice. When they are not enough, they continue to play a crucial role in avoiding data crisis and privacy breaches.

- Anonymisation techniques

Most authors are sceptical of the possibility of completely anonymising data while the original clinical record continues to exist, as individual records can be re-identified using various da-anonymisation attacks (Dreiseitl, Vinterbo & Ohno-Machado, 2001). Nevertheless, it is recognised that anonymisation should be as robust as possible such that the extract in isolation from the original clinical record would be reasonably de-identified (Fernández-Alemán *et al*, 2013). El Emam *et al.* (2013) carried out an evaluation of the risk of patient re-identification as a result of the divulging of adverse drug events reports. In doing so, they provide with an example of how such degree anonymisation can be achieved in the field of health.

A blueprint for the anonymisation of health data is broken down below in order to further aid the understanding of how the de-identification process can be applied to the particular field of health. The first step in the anonymisation process consists of the removal of direct identifiers. Data such as names, phone numbers or any other type of data that could directly identify an individual should be removed from the dataset. This is also necessary for legal compliance purposes, as it has been shown before.

Secondly, pseudo-identifiers (also known as quasi-identifiers) need to be identified. These data do not identify an individual by themselves but could if they were to appear in other datasets where they were linked to identifiers. Besides, pseudo-identifiers could enable the attacker to link records present in the same dataset. It must be noted that a classification of pseudo-identifiers does not exist. In fact, they can vary with time according to the developments in the state of the art. Some common ones in the field of health are genomic data, genealogy trees, the patient's gender or the list of interventions. Other data can help an attacker infer information which could in turn rebuild a pseudo-identifier. For instance, a certain medicine can help infer the disease it is treating. This type of data must also be identified.

Based on the identification of the previous information, the data is to be anonymised consequently. In the case of health data, randomisation, suppression and generalisation can be of great use when de-identifying datasets.

Finally, the dataset's utility and the risks that it constitutes for privacy must be assessed. While this evaluation should always take place before the publication of the datasets, it must be iterated over time to ensure that the risks remain acceptable in the presence of new datasets

and de-anonymisation techniques. Scaiano *et al*. (2016) put forward a framework for measuring the risk of re-identification in the context of unstructured medical text data releases. In their study, the traditional framework for determining the risk of re-identification and the suggested one are applied on the same dataset from the University of Michigan. The results show that conventional frameworks underestimate the risk of re-identification, which can be fixed by accounting for the context and by revising certain assumptions made by traditional approaches.

Geolocation data

The amount of geolocalised data we produce is increasing every day due to the increasing use of wearable technologies (smartphones, activity trackers, smart watches, etc.), but also because of the geolocated data produced by the vehicles we use (cars, buses, trains, etc.) or the devices that track us in public and private areas (CCTV, IMSI-catchers, Internet of Things devices, etc.). Geolocation data can be a treasure trove for a wide range of actors: researchers working on human behaviour, administrations developing a public transport network, car insurance companies, marketing companies, etc. The added value research using geolocation can bring to society is quite important. Because this data provides insights into individuals lives and behaviours though, it is essential for it to be properly handled. Anonymisation plays a central role in this data management, release and sharing, and its improper application, leading to the re-identification of individuals, can seriously harm the privacy of individuals and lead to discrimination (O'Neil, 2016).

Cases of such improper handling of geolocated data abound. One of the most salient cases of re-identification, along with breach of privacy in this context involved the New York City (NYC) taxis and seriously harmed the privacy of the taxi drivers and passengers. The NYC Taxi and Limousine Commission, following a Freedom of Information Act request, released a dataset in 2013 containing data on 173 million individual taxi trips. The pick-up and drop-off locations and times, the distance, duration, fare and tip for each record was also included. This led to many breaches of privacy. The first one concerned passengers. A lot of information could be revealed by combining known information on an individual with the dataset. A boss could check whether their employee on a business trip had correctly reported their expenses. A person could check whether their partner was lying about their whereabouts. A nosy neighbour could discover another neighbour's naughty hobbies. The home address of celebrities, and their tipping pattern could also be inferred (Tockar, 2014). The second breach concerned the taxi drivers themselves. In order to protect their identity, the taxi cab license plates and drivers' license numbers were pseudonymised. However, the regular pattern of these identifiers and the recurrence of pseudonyms across the dataset lead attackers and researchers to find that a simple algorithm had been used to pseudonymise the identifiers. They reverse-engineered the algorithm and revealed the original data, including drivers' home addresses and incomes (Douriez *et al.*, 2016). Regarding this case, Lubarsky (2017) argued that instead of applying a predetermined algorithm, the pseudonyms should have been assigned randomly.

Nevertheless, the size and one-year time-frame the dataset covered would have rendered the pseudonyms less effective anyway, due to the increased possibility of linkage attack.

The NYC taxi case led to severe breaches of privacy for thousands of individuals. Part of the problem was the fact that the NYC Taxi and Limousine Commission overlooked the risks of linkage attacks. And they are not alone -a software engineer showed that the publicly available Transport for London dataset of bike records could lead to the re-identification of individuals with the knowledge of very little outside information (Siddle, 2014).

More recently, fitness apps Strava[15] and Polar[16] also went through data protection scandals. The two track the location of their users and released, in 2017 and 2018 respectively, heat maps of their users' jogging patterns. The data was aggregated and shown as colour streams, so, theoretically, no individual running pattern could be singled out and re-identified. However, Strava and Polar did not take into account that their apps where popular among army soldiers and government officials. Therefore, the information released showed the running routes of British and American soldiers in certain conflict zones across the world, which in turn also revealed the existence of secret military bases and confidential missions.

Though it could be argued that the level of aggregation prevented individual re-identification, the information exposed could signal targets for attacks to enemy forces, thus presenting a threat for national security and putting the lives of soldiers and other state officials at risk. In this case, group-reidentification was even more serious than individual re-identification. The issues with these maps, however, was not limited to military operations. For regular users, information as sensitive as their place of residence could also be found out. In a relatively isolated zone for instance, a flow starting and ending near a cluster of homes could reveal the home address of an individual. A repeated action, in this case a running route, could both re-identify an individual and reveal personal information about them (Martijn *et al.,* 2018).

The Strava and Polar case show that aggregated data does not necessarily prevent re-identification or sensitive data from being revealed, because individuals follow patterns of behaviour across time. In fact, a study by Montjoye *et al.* (2013) showed that only 4 time-stamped location points sufficed to re-identify 97% of the million and a half individuals in the database they analysed.

These cases point to the many shortcomings weak anonymisation approaches present. As such, anonymising data may not be enough to protect people's privacy and security, especially

[15] Sly, L. (2017). U.S. soldiers are revealing sensitive and dangerous information by jogging. *The Washington Post*. Retrieved from https://www.washingtonpost.com/world/a-map-showing-the-users-of-fitness-devices-lets-the-world-see-where-us-soldiers-are-and-what-they-are-doing/2018/01/28/86915662-0441-11e8-aa61-f3391373867e_story.html?utm_term=.8e33dc26ea38 [Accessed 28 Aug. 2018].

[16] Pettit, H. (2018). Shocking security lapse as running app Polar Flow exposes the locations and personal details of 6,400 spies and personnel at MI6, the White House and GCHQ. *The Daily Mail*. Retrieved from http://www.dailymail.co.uk/sciencetech/article-5932965/Shocking-security-lapse-running-app-Polar-exposes-locations-personnel-MI6-GCHQ.html [Accessed 28 Aug. 2018].

as the risk and consequences of re-identification grow. Other safeguards should then be considered, such as limiting the access to the data and protecting with encryption mechanisms.

- Geolocation data in Data Protection legislation

An important piece of legislation at the European level concerning location data is Directive 2002/58/EC[17] on privacy and electronic communications (also known as the e-Privacy directive, which will soon be replaced by Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications).[18] The e-Privacy Directive notes that Directive 95/46/EC (the former directive which has been replaced by the GDPR) is to be applied in every instance where fundamental rights and freedoms are concerned and not specifically addressed in the e-Privacy Directive.

Location data is defined in article 2 of the e-Privacy Directive as ‘*’any data processed in an electronic communications network, indicating the geographic position of the terminal equipment of a user of a publicly available electronic communications service”*. The General Data Protection Regulation does not describe the concept of location data, but it includes it in the category of ‘’personal data’’ (article 4 GDPR). The legal regime of location data is contained in article 9 of the e-Privacy Directive, which states that location data “may be processed when they are made anonymous, or with the consent of the users or subscribers”. An interesting point to note here is that the requirement to anonymise data for processing is lifted if the consent of the data subject is given.

Similarly to GDPR, the e-Privacy directive emphasises some basic principles to bear in mind when handling location data, namely *purpose limitation* (the data will only be processed to the extent required to provide ‘’added value’’, defined in article 2 of the Directive 2002/58/EC as well), *consent* (the data subject must provide their consent in order for the processing to be lawful and can withdraw that consent at any time) and *notification* (the data subject must be notified of the processing of their data before consenting to such processing). In essence, the Directive follows the same principles that inform the General Data Protection Regulation but adapting them to the specificities of the electronic communications sector.

For mobility data, there isn’t yet a clear legal guidance or framework for data processes. There is a large array of laws regarding transport at the European Union level, and the policies on transport, one of the EU’s most strategic common policies, are mostly based on a 2011

[17] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). Retrieved from: https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:32002L0058 [Accessed 22 Aug. 2018].

[18] European Commission. Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications) (2017). Brussels. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-privacy-and-electronic-communications [Accessed 22 Aug. 2018].

European Commission White Paper creating a roadmap to a Single European Transport Area.[19] The paper only refers to personal data and privacy in paragraph 47, where it states that the protection of privacy and personal data will have to develop in parallel with the wider use of information technology tools.

- **Protecting geolocation data**

Anonymisation techniques that are applied to geolocated data are, by definition, intended to protect what is known as ''location privacy'', which can be defined as ''the ability to prevent other parties from learning one's current or past location'' (Beresford & Stajano, 2003). ''Computational location privacy'' is another way to refer to the concept and means ''the ways that computation can be used to both protect and compromise location data'' (Krumm, 2009). Shokri *et al*. (2009) highlight the importance of considering not only which anonymisation techniques are performed on the data, but also the very conceptualisation of location privacy. According to their study, the different notions present in the literature at the time of publication of the article weren't encompassing enough as they were not capturing location privacy completely in all cases. To solve that weakness, the authors proposed a novel distortion-based metric for measuring location privacy that is general enough to capture the notion of privacy under different types of location privacy mechanisms.

Duckham & Kulik (2006) classified the methods directed at the prevention of location privacy attacks into the following categories:

- Regulatory strategies: These strategies involve the development of rules to govern the fair use of personal information.
- Privacy policies: They are based on trust, not on coercion. They enable the tailoring of the privacy requirements to the needs of individuals and specific transactions.
- Anonymity: These solutions are based around the dissociation between the user's identity and information about them.
- Obfuscation: It implies a process of degrading the quality of the information about a person's location with the aim of protecting their privacy while keeping utility.

This last approach, obfuscation, is a form of *pseudonymisation* that tries to optimally combine privacy with utility. However, obfuscation techniques have raised questions regarding their efficacy, since de-anonymisation attacks have proven to be successful given the far from random character of the movement patterns of people, which makes them easily identifiable as we saw earlier (Gambs, Killijian & del Prado Cortez, 2010, 2014).

[19] European Commission (2011) Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system [White paper] Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52011DC0144&from=EN [Accessed 22 Aug. 2018].

Below we address the methods and techniques that can be used to anonymise geolocation data in three different contexts: geolocation apps, biomedical data and mobility data. This is because these three areas present specific characteristics, but also because they are very relevant in our daily activities in different ways. Firstly, most of the apps we use in our phones gather geolocated data, regardless of whether this is essential for the provision of the service they offer or not, and many scientific processes based on crowdsourcing of data also collect geolocated data. Secondly, many wearables capture geolocation data for health purposes. Therefore, protecting privacy in geolocated medical data is of utmost importance, and we address it specifically and as a way to complement the previous section on health data. Thirdly, geolocation data is of particular importance in mobility services, as it allows for the improvement of routes and services. In this latter case, the data collected may not be directly personal, as in the case of taxi or bus use data but could easily lead to re-identification.

- In geolocation apps

In 2013 Andres *et al*. suggested an anonymisation method which builds upon *differential privacy*, but that also incorporates a mechanism to achieve "geo-indistinguishability''. This is attained by adding controlled random noise to the user's location. According to the authors, this system would improve privacy on LBS (Location Based Systems) applications without affecting negatively their functionality. Rastogui and Nath (2010) also put forward algorithms with similar aims that are more complex since they also account for correlations.

A different type of application involving the use of geolocated data are those functioning through ''crowdsourcing'', which according to Boukoros *et al*. (2018) implies that participants collect geolocated data on their devices and share it with these application's central servers, which in turn process them to attain a particular objective. Within this group of applications, we can find, those that allow for map enrichment through geo-tagged photos or those which participants can use to suggest places or routes to other participants. In their study, the authors try to gauge the implications that sharing content through these platforms could have for privacy by focusing on three of the most popular ones (Safecast,[20] Radiocells[21] and Open Street Maps[22]). The effectiveness of three main anonymisation methods (*spatial obfuscation*, *hiding* and *generalisation*) is tested with datasets coming from the three previously mentioned apps. They conclude that the current state of the art is insufficient to address privacy threats in this domain. Moreover, regulation does not seem to be able to solve the issue, given that users give free consent to share their data under permissive licenses for altruistic ends, which leaves the development of technological solution as the only alternative.

- In biomedical research

A recent article addressed the question of privacy of mobile location data used in biomedical research (Goldenholz *et al.*, 2018). In it, the authors conclude that it is virtually impossible to

[20] Safecast. Retrieved from https://blog.safecast.org/ [Accessed 22 Aug. 2018].
[21] Radiocells.org. Retrieved from https://radiocells.org/ [Accessed 22 Aug. 2018].
[22] OpenStreetMap. Retrieved from https://www.openstreetmap.org/ [Accessed 22 Aug. 2018].

achieve complete anonymity. They argue that the techniques used must vary on a case by case basis according to the risks posed by the characteristics of the data and to the state of technology at a given moment. They list the following strategies to anonymise datasets which include mobile location:

- *Data aggregation* when the groups within the datasets are sufficiently large. To ensure these groups are in fact sufficiently large, *k-anonymity* can be used.
- *Access control*: location data is made unreadable unless granted permission for use. The authors include under this umbrella removal, encryption, noise addition, decrease in resolution, simulation and cloaking. All of these, except encryption, involve a certain loss in the quality and efficacy of the data.
- Preserve only a portion of the location data: instead of fixed position, relative distances can be given (*generalisation*) and location can be altered to become less accurate (*noise addition*).
- *Temporal manipulations*: The temporal relationship between locations could also lead to the re-identification of an individual. In this case, the same manipulation as the spatial manipulations could be performed.
- When location data can be linked to social-demographic data attributes (age, gender, ethnicity, income, etc.), the value of these attributes can be randomly *swapped* to prevent linkage. This manipulation of the attributes should only be done if the linking between the locations and the attributes is not required for further analysis.

    - In mobility services

Research done on data aggregation as an anonymisation strategy used for the publication of location datasets, which are used to improve citizen services in smart cities or by the industry to monetise location data, has also been conducted (Pyrgelis, Troncoso and De Cristofaro, 2018). This research challenges the view that "by grouping users' traces the aggregates no longer contain personal data and thus can be freely shared for various analytics tasks" (ibid, 2018, p.1). Very regular or very uncommon patterns of mobility appear to be easier to recognise and users that contributed largely to the dataset are thus easier to spot.

Pyrgelis *et al.* (2018) found that in terms of anonymisation strategies, it is feasible to build defences against re-identification attacks if the dataset will be used for particular applications. Conversely, it appears to be extremely challenging to do the same when the potential uses for the data are unspecified, as is the case in most Open Data schemes. The anonymisation technique that provided the best results in terms of utility and privacy in the study was data *generalisation* and hiding techniques (*suppression* and *sampling*). Though *spatial generalisation* (making locations less specific) did not provide significant protection against membership attacks (being able to infer whether an individual has participated in a dataset), *temporal generalisation* (making times less specific) improved privacy. The authors noted that

applying spatial and temporal generalisation simultaneously will increase privacy. Furthermore, *sampling* proves to be useful when the input signal is sparse.

An anonymisation method that has been suggested in the literature in the field of geolocation is the creation of *synthetic location traces* which can plausibly reflect individuals with consistent lifestyles and meaningful mobilities (Bindschaedler & Shokri, 2016). This type of synthetic data would allow, according to the authors, for the preservation of the utility of the data and for significant improvements in terms of preventing location inference attacks from happening, since the statistically crafted traces do not leak significant information about any particular individual whose data is used to produce the synthetic dataset. Those would be significant improvements since current techniques for generating fake locations are neither resilient against location inference attacks nor able to produce data useful for many applications. Overall however, Duckham, Kulik and other authors agree that there is currently not a single strategy capable of solving all problems posed by location tracking devices in terms of privacy. Subsequently, privacy solutions will necessarily have to combine measures coming from all four typologies. Stress is put on the fact that regulations affecting location data tend to make no distinctions between static data (name, address) and dynamic data (location), which makes it difficult to develop privacy policies with a temporal element (Duckham *et al.*, 2006).

However difficult it may be to strike a balance between anonymisation and utility, the efforts to improve current approaches continue. Uber, for instance, recently launched a project called ''Movement'', which provides the anonymised data "of over two billion trips to help urban planning around the world".[23] The platform provides tools to analyse traffic changes and patterns over time. The nature of the anonymisation method performed onto the data is briefly described in a downloadable available on their website. Travel time data are removed for zones that do not meet a minimum number of trips and do not meet the minimum count of unique drivers and riders necessary to preserve driver/rider privacy.[24] This endeavour has also been taken up by other companies such as Easy Taxi, Grab and Le Taxi that have joined a partnership with the World Bank and other organisations to make their data available through an Open Data Licence.[25]

[23] Uber Movement: Let's find smarter ways forward. (2018). Retrieved from https://movement.uber.com/?lang=es-ES [Accessed 22 Aug. 2018].
[24] Uber Movement: Travel Times Calculation Methodology. (2018). Retrieved from https://d3i4yxtzktqr9n.cloudfront.net/web-movement/static/pdfs/Movement-TravelTimesMethodology-76002ded22.pdf [Accessed 22 Aug. 2018].
[25] Ribeiro, J. (2017). Uber offers cities 'anonymized' ride data. *Pcworld*. Retrieved from https://www.pcworld.com/article/3155494/techology-business/uber-to-provide-anonymized-data-to-city-officials.html [Accessed 28 Aug. 2018].

Statistics

Most organisations collect personal data. This data is often released for research or statistical purposes for the scientific community. As such, the data has normally already undergone a process of anonymisation when it is released for statistical purposes. Data in statistics, by definition, then undergo a process of organisation, analysis and interpretation (Romeijn, 2017), thus further increasing anonymity -in theory. This assumption explains why, as mentioned below, laws pertaining to statistics only demand that statistics comply with data protection laws, and do not demand the anonymisation of statistical datasets: Anonymised datasets do not need to be anonymised, as they should already be clear of any identifiable data. Census data for instance, is anonymised data specifically collected for statistical purposes. The statistics drawn from the census should not lead to the re-identification of individuals.

In practice however, individuals can be re-identified from statistical data (Cox, 1985). For instance, in 2008 Homer *et al.* (2008), mentioned above, re-identified individuals in genome-wide association studies (GWAS), though these studies report only summary statistics on hundreds of thousands of individuals. This and other studies led the US National Institute of Health to pull GWAS from public databases and require specific permissions to be obtained to access the data (Check Hayden, 2013). This demonstrates the shortcomings of anonymisation even in the case of statistics, and the need to not only improve anonymisation, but also include other methods of responsible and ethical data management (access control, for instance) to counteract the shortcomings of anonymisation.

- Statistics in Data Protection legislation

The processing of personal data for statistical purposes is regulated by the GDPR. Anonymised data does not fall within the scope of the Regulation, hence anonymised data used for statistical purposes also falls outside of the scope of the Regulation. Statistical purposes mean “any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results” (GDPR Recital 162, 2016). Article 89 specifically addresses processing for purposes including statistical purposes. It demands for appropriate safeguards to be put in place to ensure the rights and freedoms of data subjects and to respect the principle of data minimisation (Article 89(1)) and mentions pseudonymisation as one of the possible measures. The article also allows Union or Member states to provide derogation from certain data subject rights in specific cases.

Regulation (EC) No 223/2009[26] on European statistics specifies the rules laid down in the data protection directive (now GDPR) as far as European statistics are concerned and ensures the

[26] Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and

protection of personal data (Recital 22). Article 19 addresses the question of the release of public use files, which should be disseminated ''*in the form of a public use file consisting of anonymised records which have been prepared in such a way that the statistical unit cannot be identified, either directly or indirectly, when account is taken of all relevant means that might reasonably be used by a third party*''. overall, the Regulation demands a reasonable degree of anonymisation of personal data before its dissemination with almost no exceptions (article 20 Regulation 223/2009).

A couple of considerations need to be made in this context. First, that Regulation 223/2009 on European statistics seldom mentions personal data (art. 19-20, recital 22 Regulation 223/2009), showing that statistical data is considered anonymised data. Second, that GDPR provides exceptions for the processing of statistical data, if it is shown that data protection measures would hinder innovation (art. 89 GDPR). This open the door to less robust approaches to data protections in this field. However, in the case of a breach or a data protection failure, the data controller would still need to prove that sufficient safeguards were put in place.

- **Protecting statistical data**

As statistical data should already have been anonymised, we focus here on data that goes through the process of anonymisation to become statistical data. In order to illustrate how data is processed and anonymised for statistical purposes to respect privacy rights, we draw upon the example of a survey about housing conditions conducted by the Northern Ireland Housing Executive (NIHE). Every two years, this agency carries out a survey known as the Northern Ireland House Condition Survey, in which data regarding the housing stock and its dwellers is gathered. The NIHE then publishes a report providing statistics and insights based on the analysis the data.[27] Guidelines to ensure the quality of the data released are also published. These include procedures to ensure the confidentiality of the survey before publication,[28] such as briefing of respondents concerning the privacy and confidentiality of their information, the training of staff working on the Survey, the restricted access to the information by NIHE staff, and the statistical disclosure control (SDC) techniques used to ensure no individual can be identified once the statistics are released. SDCs are used to ensure that no individual, business

Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities (Text with relevance for the EEA and for Switzerland). Retrieved from: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32009R0223 [Accessed 22 Aug. 2018].

[27] House Condition Survey Main Report | The Housing Executive (Touch). Retrieved from https://touch.nihe.gov.uk/house_condition_survey_main_report_2016.pdf [Accessed 22 Aug. 2018].

[28] Quality Information | The Housing Executive (Touch). (2018). Retrieved from https://touch.nihe.gov.uk/index/corporate/housing_research/house_condition_survey/corporate-quality-information.htm [Accessed 22 Aug. 2018].

or other subject of study are identifiable from published data, and include rounding, aggregation and suppression.[29]

The UK Office for National Statistics has published various guides for disclosure control, including for "social survey microdata" (such as the data from the NIHE survey). The methods they put forward are the following:[30]

- *Aggregation*: some key variables can be recoded so there are fewer categories. Ages can be recorded in 5- or 10-year groups, apply top-bottom coding to annual incomes so that the highest category is '£100,000 and above', make the geographical information less precise, etc.
- *Perturbation*: add a random amount to certain reported values, change the categories of some rare values
- *Suppression*: suppress the unique or rare records, or records with unique or rare combinations of attributes. For instance, already in the 1980s, the census bureau would not disclose certain records if less than 15 people or 5 households shared the same attributes (Lawrence H. Cox 1985). The guide warns against the damage such a measure could cause in terms of analysis.
- *Removal* of key variables: if some variables are particularly disclosive, an option would be to remove them all-together.
- *Swap* records: if records in different geographical locations present the same key variables, they can be swapped.

In the case of health, the Office for National Statistics gives the same guidelines, though presented differently.[31]

Releasing *synthetic data* is also a possible option, which can counter the distortion in insights that the previously mentioned techniques create. In 1993 already, Rubin proposed the release of synthetic datasets to the public. Since then, many have promoted the use of synthetic datasets (Reither, 2005; Raghunathan Reiter & Rubin, 2003) as a way to provide valid inferences and analysis from the data. Although Rubin advocated for the use of a fully synthetic datasets, others propose the disclosure of only partially synthetic datasets, with only the most

[29] Northern Ireland Housing Executive. (2017). *Confidentiality and Access Statement*. Retrieved from: https://www.nihe.gov.uk/confidentiality_and_access_statement.pdf [Accessed 22 Aug. 2018].
[30] Office for National Statistics. (2014). GSS/GSR Disclosure Control Guidance for Microdata Produced from Social Surveys. Retrieved from: https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyforsocialsurveymicrodata [Accessed 22 Aug. 2018].
[31] Office for National Statistics. (2006). Review of the Dissemination of Health Statistics: Confidentiality Guidance. Retrieved from: https://www.ons.gov.uk/file?uri=/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/healthstatistics/confidentialityguidanctcm77181864.pdf [Accessed 22 Aug. 2018].

sensitive attributes being synthesised. This can render the data more accurate, and thus more helpful (Reiter, 2005).

The steps taken to decide on which anonymisation method to use are the same as with health data: identifiers, quasi-identifiers and other types of data must be identified before the anonymisation, and risks of re-identification must be assessed after.

In conclusion, agencies that manage personal data and process it in order to produce useful statistical information must enact measures to ensure that legal and ethical standards related to data protection and privacy are met. To do so, using a combination of anonymisation methods and governance schemes is necessary. When it comes to anonymisation techniques, the choice will always be context-dependent.

Bibliography

Agrawal, R., & Johnson, C. (2007). Securing electronic health records without impeding the flow of information. *International Journal of Medical Informatics*, *76*(5-6), 471-479. doi: 10.1016/j.ijmedinf.2006.09.015

Authority, U. S. (2009). Code of practice for official statistics. *Retrieved from: statisticsauthority.gov.uk/assessment/code-of-practice* [Accessed 22 Aug. 2018].

Beresford, A. R., & Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive computing*, (1), 46-55. Retrieved from: https://www.utdallas.edu/~muratk/courses/privacy08f_files/location_privacy_pervasive_computing.pdf [Accessed 22 Aug. 2018].

Beresford, A. R., & Stajano, F. (2004, March). Mix zones: User privacy in location-aware services. In *Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on* (pp. 127-131). IEEE. Retrieved from: https://www.cl.cam.ac.uk/research/dtg/www/files/publications/public/arb33/BeresfordStajano-MixZones-PerSec2004.pdf [Accessed 22 Aug. 2018].

Bindschaedler, V., & Shokri, R. (2016, May). Synthesizing plausible privacy-preserving location traces. In *Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 546-563). IEEE.

Boukoros, S, Humbert, M., Katzebeisser, S., Troncoso, C. Enhancing Privacy in the Wild: The Case of Mobile Crowdsourcing Applications. 2018.

Check Hayden, E. (2013). Privacy protections: The genome hacker. Nature, 497(7448). Retrieved from https://www.nature.com/news/privacy-protections-the-genome-hacker-1.12940#/ref-link-4 [Accessed 28 Aug. 2018].

Corti, L., Day, A., & Backhouse, G. (2000, December). Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* (Vol. 1, No. 3). Retrieved from: http://www.qualitative-research.net/index.php/fqs/article/download/1024/2208 [Accessed 22 Aug. 2018].

Cox, L. H., Johnson, B., McDonald, S. K., Nelson, D., & Vazquez, V. (1985, March). Confidentiality issues at the Census Bureau. In *Proceedings of the First Annual Census Bureau Research Conference, Washington, DC: US Government Printing Office* (pp.199-218).

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*, 1376. Retrieved from: https://www.nature.com/articles/srep01376?ial=1 [Accessed 22 Aug. 2018].

Douriez, M., Doraiswamy, H., Freire, J., & Silva, C. T. (2016, October). Anonymizing NYC Taxi Data: Does It Matter?. In *Data Science and Advanced Analytics (DSAA)*, 2016 IEEE International Conference on (pp. 140-148). IEEE.

Dreiseitl, S., Vinterbo, S., & Ohno-Machado, L. (2001). Disambiguation data: extracting information from anonymized sources. In *Proceedings of the AMIA Symposium* (p. 144). American Medical Informatics Association. Retrieved from:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243291/pdf/procamiasymp00002-0183.pdf [Accessed 22 Aug. 2018].

Duckham, M., & Kulik, L. (2005, May). A formal model of obfuscation and negotiation for location privacy. In *International conference on pervasive computing* (pp. 152-170). Springer, Berlin, Heidelberg. Retrieved from: http://profs.sci.univr.it/~giaco/download/Watermarking-Obfuscation/obfuscation%20locations.pdf [Accessed 22 Aug. 2018].

Duckham, M., & Kulik, L. (2006). Location privacy and location-aware computing. In *Dynamic and Mobile GIS* (pp. 63-80). CRC Press. Retrieved from: https://www.researchgate.net/profile/Lars_Kulik/publication/266408992_Location_privacy_and_location-aware_computing/links/54cee5050cf29ca810fd0e7f/Location-privacy-and-location-aware-computing.pdf [Accessed 22 Aug. 2018].

El Emam, K., Dankar, F., Neisa, A., & Jonker, E. (2013). Evaluating the risk of patient re-identification from adverse drug event reports. *BMC Medical Informatics And Decision Making*, *13*(1). doi: 10.1186/1472-6947-13-114 Retrieved from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-13-114 [Accessed 22 Aug. 2018].

Fernández-Alemán, J. L., Señor, I. C., Lozoya, P. Á. O., & Toval, A. (2013). Security and privacy in electronic health records: A systematic literature review. *Journal of biomedical informatics*, *46*(3), 541-562. Retrieved from: https://www.sciencedirect.com/science/article/pii/S1532046412001864 [Accessed 22 Aug. 2018].

Ferraiolo, D., Cugini, J., & Kuhn, D. R. (1995, December). Role-based access control (RBAC): Features and motivations. In *Proceedings of 11th annual computer security application conference* (pp. 241-48). Retrieved from: https://www.researchgate.net/profile/D_Kuhn2/publication/238743515_Role-based_access_control_features_and_motivations/links/562e7d3808ae04c2aeb5d98b.pdf [Accessed 22 Aug. 2018].

Ferraiolo, D.F. & Kuhn, D.R. (October 1992). "Role-Based Access Control". 15th National Computer Security Conference: 554–563. Retrieved from: http://csrc.nist.gov/groups/SNS/rbac/documents/ferraiolo-kuhn-92.pdf [Accessed 22 Aug. 2018].

Finkenzeller, K. (2010). *RFID handbook: fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication*. John Wiley & Sons. Retrieved from: http://117.3.71.125:8080/dspace/bitstream/DHKTDN/6830/1/6183.RFID%20Handbook%20%283rd%20ed%29.pdf [Accessed 22 Aug. 2018].

Gambs, S., Killijian, M. O., & del Prado Cortez, M. N. (2010, November). Show me how you move and I will tell you who you are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS* (pp. 34-41). ACM. Retrieved from: https://www.researchgate.net/profile/Marc-Olivier_Killijian/publication/221589955_Show_Me_How_You_Move_and_I_Will_Tell_You_Who_You_Are/links/0c96051a614a4356b5000000.pdf [Accessed 22 Aug. 2018].

Gambs, S., Killijian, M. O., & del Prado Cortez, M. N. (2014). De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, *80*(8), 1597-1614. Retrieved from: https://hal.archives-ouvertes.fr/hal-01242268/document [Accessed 22 Aug. 2018].

Gedik, B., & Liu, L. (2008). Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *IEEE Transactions on Mobile Computing*, *7*(1), 1-18. Retrieved from: https://www.researchgate.net/profile/Bugra_Gedik/publication/3436752_Liu_L_Protecting_Location_Privacy_with_Personalized_k-Anonymity_Architecture_and_Algorithms_IEEE_Transactions_on_Mobile_Computing_71_1-18/links/00b7d5326f2880613c000000/Liu-L-Protecting-Location-Privacy-with-Personalized-k-Anonymity-Architecture-and-Algorithms-IEEE-Transactions-on-Mobile-Computing-71-1-18.pdf [Accessed 22 Aug. 2018].

Goldenholz, D. M., Goldenholz, S. R., Krishnamurthy, K. B., Halamka, J., Karp, B., Tyburski, M., ... & Theodore, W. (2018). Using mobile location data in biomedical research while preserving privacy. *Journal of the American Medical Informatics Association*.

Guide, Q. P. (2002). Qualitative Data Processing. *UK Data Archive*. Retrieved from: https://sp.ukdataservice.ac.uk/qualidata/documents/dataprocess.pdf [Accessed 22 Aug. 2018].

Gulcher, J. R., Kristjánsson, K., Gudbjartsson, H., & Stefánsson, K. (2000). Protection of privacy by third-party encryption in genetic research in Iceland. *European journal of human genetics*, *8*(10), 739.

Hammer, C., Kostroch, M. D. C., & Quiros, M. G. (2017). *Big Data: Potential, Challenges and Statistical Implications*. International Monetary Fund. Retrieved from: https://www.imf.org/~/media/Files/Publications/SDN/2017/sdn1706-bigdata.ashx [Accessed 22 Aug. 2018].

Hartzog, W., & Rubinstein, I. (2017). The anonymization debate should be about risk, not perfection. Communications of the ACM, 60(5), 22-24.

Hassan, W. U., Hussain, S., & Bates, A. (2018, August). Analysis of Privacy Protections in Fitness Tracking Social Networks-or-You can run, but can you hide?. In 27th {USENIX} Security Symposium ({USENIX} Security 18) (pp. 497-512). USENIX} Association}. Retrieved from: https://www.usenix.org/sites/default/files/conference/protected-files/security18_slides_hassan.pdf [Accessed 22 Aug. 2018].

Hay M., Machanavajjhala, A., Miklau, G., Chen, Y., and Zhang, D. 2016. Principled Evaluation of Differentially Private Algorithms using DPBench. In Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16). ACM, New York, NY, USA, 139-154.

Henrici, D., & Muller, P. (2004, March). Hash-based enhancement of location privacy for radio-frequency identification devices using varying identifiers. In Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on (pp. 149-153). IEEE. Retrieved from: http://dspace.icsy.de:12000/dspace/bitstream/123456789/124/1/DPArchiv.0080.pdf [Accessed 22 Aug. 2018].

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., & Muehling, J. et al. (2008). Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. Plos Genetics, 4(8), e1000167. doi: 10.1371/journal.pgen.1000167

Humbert, M., Ayday, E., Hubaux, J., & Telenti, A. (2013). Addressing the concerns of the lacks family: Quantification of kin genomic privacy. Conference On Computer And Communications Security

(CCS). Retrieved from:https://infoscience.epfl.ch/record/188347/files/fp320-humbertAemb.pdf [Accessed 22 Aug. 2018].

Iacono, L. L. (2007). Multi-centric universal pseudonymisation for secondary use of the EHR. *Studies in health technology and informatics*, *126*, 239. Retrieved from: https://www.researchgate.net/profile/Luigi_Lo_Iacono2/publication/6354273_Multi-centric_Universal_Pseudonymisation_for_Secondary_Use_of_the_EHR/links/02e7e51bf0f78c9c0d000000.pdf [Accessed 22 Aug. 2018].

Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of Big Data*, *3*(1), 25. Retrieved from: https://link.springer.com/content/pdf/10.1186%2Fs40537-016-0059-y.pdf [Accessed 22 Aug. 2018].

Juels, A. (2006). RFID security and privacy: A research survey. *IEEE journal on selected areas in communications*, *24*(2), 381-394. Retrieved from: https://www.researchgate.net/profile/Ari_Juels/publication/3236246_RFID_security_and_privacy_A_research_survey/links/00b4953bbe80a8c975000000/RFID-security-and-privacy-A-research-survey.pdf [Accessed 22 Aug. 2018].

Keßler, C., & McKenzie, G. (2018). A geoprivacy manifesto. *Transactions in GIS*, *22*(1), 3-19. Retrieved from: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12305 [Accessed 22 Aug. 2018].

Kitchin, R. (2015). Big data and official statistics: Opportunities, challenges and risks. Retrieved from: http://eprints.maynoothuniversity.ie/7231/1/PC [Accessed 22 Aug. 2018].

Krumm, J. (2009). A survey of computational location privacy. *Personal and Ubiquitous Computing*, *13*(6), 391-399. Retrieved from: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/12/computational-location-privacy-preprint.pdf [Accessed 22 Aug. 2018].

Kushida, C. A., Nichols, D. A., Jadrnicek, R., Miller, R., Walsh, J. K., & Griffin, K. (2012). Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, S82-S101. Retrieved from: https://s3.amazonaws.com/academia.edu.documents/42733508/Strategies_for_de-identification_and_ano20160216-21787-61su1m.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1533826528&Signature=XeUENA8S1yosUqJuvll508Q1nYY%3D&response-content-disposition=inline%3B%20filename%3DStrategies_for_de-identification_and_ano.pdf [Accessed 22 Aug. 2018].

Lubarsky, B. (2017). *Re-identification of "Anonymized Data."* Georgetown Law Technology Review, 202.

Martijn, M., Tokmetzis, D., Bol, R., & Postma, F. (2018). This fitness app lets anyone find names and addresses for thousands of soldiers and secret agents. *De Correspondent*. Retrieved from https://decorrespondent.nl/8480/this-fitness-app-lets-anyone-find-names-and-addresses-for-thousands-of-soldiers-and-secret-agents/260810880-cc840165 [Accessed 22 Aug. 2018].

Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: differential privacy for location-based systems. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security (CCS '13). ACM, New York, NY, USA, 901-914

Mokbel, M. F., Chow, C. Y., & Aref, W. G. (2006, September). The new casper: Query processing for location services without compromising privacy. In *Proceedings of the 32nd international conference on Very large data bases* (pp. 763-774). VLDB Endowment. Retrieved from: https://infolab.usc.edu/csci599/Fall2009/papers/The%20New%20Casper%20Query%20Processing%20for%20Location%20Services%20without%20Compromising%20Privacy.pdf [Accessed 22 Aug. 2018].

Moreno, J., Serrano, M. A., & Fernández-Medina, E. (2016). Main issues in big data security. *Future Internet*, *8*(3), 44. Retrieved from: https://www.mdpi.com/1999-5903/8/3/44/html [Accessed 22 Aug. 2018].

Neubauer, T., & Heurix, J. (2011). A methodology for the pseudonymization of medical data. *International journal of medical informatics*, *80*(3), 190-204.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. Ucla L. Rev., 57, 1701. Retrieved from: https://pages.uoregon.edu/koopman/courses_readings/phil407-net/ohm_broken_promises_privacy.pdf [Accessed 22 Aug. 2018].

O'Neil, C. (2016). Weapons of Math Destruction. [S.I.]: Crown/Archetype.

Privacy Analytics (2017). European legal requirements for use of anonymized health data for research purposes by a data controller with access to the original (identified) datasets. Retrieved from: https://iapp.org/resources/article/european-legal-requirements-for-use-of-anonymized-health-data-for-research-purposes-by-a-data-controller-with-access-to-the-original-identified-data-sets/ [Accessed 22 Aug. 2018].

Pyrgelis A., Troncoso, C. and De Cristofaro E. Unraveling the Mysteries of Membership Inference. Manuscript submitted for publication.

Pyrgelis, A., Troncoso, C., & De Cristofaro, E. (2017). Knock Knock, Who's There? Membership Inference on Aggregate Location Data. *arXiv preprint arXiv:1708.06145*.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, *19*(1), 1. Retrieved from: http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/multiple-imputation-for-statistical-disclosure-limitation.pdf [Accessed 22 Aug. 2018].

Rastogi V. and Nath S. Differentially private aggregation of distributed time-series with transformation and encryption. In SIGMOD, 2010

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, *18*(4), 531. Retrieved from: http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/satisfying-disclosure-restrictions-with-synthetic-data-sets.pdf [Accessed 22 Aug. 2018].

Reiter, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(1), 185-205. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.3879&rep=rep1&type=pdf [Accessed 22 Aug. 2018].

Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, *21*(3), 441. http://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/using-cart-to-generate-partially-synthetic-public-use-microdata.pdf [Accessed 22 Aug. 2018].

Romeijn, J. (2017). Philosophy of Statistics. Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/entries/statistics/ [Accessed 28 Aug. 2018].

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, *9*(2), 461-468. Retrieved from: http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf [Accessed 22 Aug. 2018].

Sax, U., Kohane, I., & Mandl, K. D. (2005). Wireless technology infrastructures for authentication of patients: PKI that rings. *Journal of the American Medical Informatics Association*, *12*(3), 263-268. Retrieved from: https://watermark.silverchair.com/12-3-263.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAaUwggGhBgkqhkiG9w0BBwagggGSMIIBjgIBADCCAYcGCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMidwJaIygpzZx6DbFAgEQgIIBWNrrikbVuCV4Qf6D0LK2jUpMa1EXdQXUXoKw-RiyQqtDgxeFxSgQEC-uP8kxsk1ITdwZtK4u3k7tu9FTDlxkgk4vZ20MJSZyOFuxCJgNhovMhVRMNzc_gZgTkS2SYSjLJGNhPJw7ONcRotcNKVKq_0Vic6ieL0_GCDwmPCcuK_Xtc2oLcJlO6ygW239NfjQFNdC39oUlktZFPLwMDiDEkv29HEm1KTr-mnMEoKCm06-YnfBjBp_Rsn_eG8ivfuBvjzTnYRZ_VF0QB3Pk05nQilC8KoOOchxdWhptsKM4b3HdBH-dAQw_qQH0BX5PbrwNfLIxa4VC0tosVa0dxM2jF0vNsqSOJAU2ZF629X_1ZEIyaYHI8_DZZjC4A1i_uP3CfvydUIVlHE3WtrzXXi5vxnZ5dVkEVnHbRkrFr1aAE7FzvhZw0iSMMheSLWaTbgzdHn3J5g3LjM8g [Accessed 22 Aug. 2018].

Scaiano, M., Middleton, G., Arbuckle, L., Kolhatkar, V., Peyton, L., Dowling, M., ... & El Emam, K. (2016). A unified framework for evaluating the risk of re-identification of text de-identification tools. *Journal of biomedical informatics*, *63*, 174-183.

Shokri, R., Freudiger, J., Jadliwala, M., & Hubaux, J. P. (2009, November). A distortion-based metric for location privacy. In *Proceedings of the 8th ACM workshop on Privacy in the electronic society* (pp. 21-30). ACM.

Shringarpure, S., & Bustamante, C. (2015). Privacy Risks from Genomic Data-Sharing Beacons. The American Journal Of Human Genetics, 97(5), 631-646. doi: 10.1016/j.ajhg.2015.09.010

Siddle, J. (2014). I Know Where You Were Last Summer: London's public bike data is telling everyone where you've been [Blog]. Retrieved from https://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html [Accessed 22 Aug. 2018].

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco)*, *671*, 1-34. Retrieved from: http://ggs685.pbworks.com/w/file/fetch/94376315/Latanya.pdf [Accessed 22 Aug. 2018].

Thomson, D., Bzdel, L., Golden-Biddle, K., Reay, T., & Estabrooks, C. A. (2005, January). Central questions of anonymization: A case study of secondary use of qualitative data. In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research (Vol. 6, No. 1). Retrieved from: http://www.qualitative-research.net/index.php/fqs/article/viewFile/511/1103.. [Accessed 22 Aug. 2018].

Tockar, A. (2014). Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset. Retrieved from https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/ [Accessed 22 Aug. 2018].

Wang, R., Li, Y., Wang, X., Tang, H., & Zhou, X. (2009). Learning your identity and disease from research papers: information leaks in genome wide association study. Proceedings Of The 16Th ACM Conference On Computer And Communications Security. Retrieved from https://www.cs.indiana.edu/pub/techreports/TR680.pdf [Accessed 22 Aug. 2018].

Willenborg, L., & De Waal, T. (1996). *Statistical disclosure control in practice* (Vol. 111). Springer Science & Business Media.

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: privacy and data mining. *IEEE Access*, *2*, 1149-1176. Retrieved from: https://ieeexplore.ieee.org/iel7/6287639/6705689/06919256.pdf [Accessed 22 Aug. 2018].