Received: December 17, 2019, Date of acceptance: March 06, 2020, Date of publication: April 16, 2020, DOI: https://doi.org/ 10.26826/law-in-context.v36i2.110

Misconceptions in Privacy Protection and Regulation

By **Chris Culnane** (Honorary Fellow in the School of Computing and Information Systems at the University of Melbourne), **Orcid**: https://orcid.org/0000-0002-9543-1342
and Kobi Leins (Senior Research Fellow in Digital Ethics with School of Computing and Information Systems at the University of Melbourne and Non-Resident Fellow with the United Nations Institute of Disarmament Research), **Orcid**: https://orcid.org/0000-0002-9432-5724
University of Melbourne, Australia

ABSTRACT

Privacy protection legislation and policy is heavily dependent on the notion of de-identification. Repeated examples of its failure in real-world use have had little impact on the popularity of its usage in policy and legislation. In this paper we will examine some of the misconceptions that have occurred to attempt to explain why, in spite of all the evidence, we continue to rely on a technique that has been shown not to work, and further, which is purported to protect privacy when it clearly does not. With a particular focus on Australia, we shall look at how misconceptions regarding de-identification are perpetuated. We highlight that continuing to discuss the fiction of de-identified data as a form of privacy actively undermines privacy and privacy norms. Further, we note that 'de-identification of data' should not be presented as a form of privacy protection by policy makers, and that greater legislative protections of privacy are urgently needed given the volumes of data being collected, connected and mined.

Keywords – *privacy, re-identification, de-identification, anonymity, policy*

Acknowledgements. *We would like to thank the anonymous reviewers for their detailed and insightful comments.*

Disclosure statement – *No potential conflict of interest was reported by the author.*

Suggested citation: Chris Culnane and Kobi Leins, "Misconceptions in Privacy Protection and Regulation" *Law in Context*, 36 (2): 49-60. DOI: https://doi.org/10.26826/law-in-context.v36i2.110

Summary

ISSN: 1839-4183

1. INTRODUCTION

We exist in a time of unprecedented data collection, processing, and sharing, creating what Zuboff refer to as "surveillance capitalism". Current business practices rely on constant collection of all kinds of data, often collected without consent or event knowledge of the collection. It is not without reason that the terminology that surrounds this data activity remains contested.

The purpose of the *Privacy Act 1988* (Cth) (hereinafter the *Privacy Act*) is to protect privacy and prevent the collection and use of certain types of information. This raises the question, how can we have a Privacy Act that is supposed to provide protections, yet those protections appear not to materialise in practice? Zuboff notes that "[privacy] is not eroded but redistributed, as decision rights over privacy are claimed for surveillance capital" (Zuboff 2019, p. 90).

In this paper we will examine how and why privacy law has been so profoundly eroded, and in particular, identify the role of de-identification as a method for by-passing the protections provided by the Privacy Act. We will examine how de-identification is inaccurately defined, and how this results in large-scale exploitation of data by industry. What Zuboff coins "surveillance capitalism", she argues, is far too profitable and unregulated to genuinely contemplate human rights, civil liberties or public benefit (Zuboff 2019). Longer term, unclear or even misleading definitions, alongside inadequate policy approaches, are being embedded in policy discourse, directly undermining the Privacy Act, and further eroding privacy through recent proposals around data sharing and use.

2. DEFINITIONS OF DATA

One of the challenges in regulating and legislating to protect privacy is that the protection is heavily dependent on technical definitions of data and its use. The rapid rate of change in technology has led to a level of dynamism in such definitions, which in turn has led to ambiguity, inconsistency, and even contradictory definitions of the same term. In turn, terms such as "data exhaust" imply that the data does not require any privacy protections, as it has no real value. In fact, nothing could be further from the truth (Zuboff 2019, p. 68). In addition to the value of what is frequently referred to as data detritus, the ability to connect seemingly irrelevant datapoints to create a source of information far bigger than 'the sum of its parts' creates challenges of its own, not least of which because "identifiability" becomes easier with the increased frequency with which such data is collected.

To give a more concrete example of one type of data that may be collected, let's take the example of clicking on an article online. The host of the platform, as well as their third-party advertisers, can tell what kind of device you're on, what browser you use, what you do on the site (what articles you read, how long you stay on each, what ads you click on) and what site you visit next when you click somewhere else, as well as where you are based on your device's unique IP address (Moller 2019). None of these data points contain traditional identifiers, but they are a unique fingerprint for your device[1], and by extension you – particularly when using a mobile device, which tends to be a single user device. As such, a detailed profile of your preferences, actions, and history can be constructed, all seemingly without ever having to know your name. Whenever you visit a future site, the same unique signature can be used to look up your profile and display targeted adverts, or dynamically price goods and services.

2.1. Personal Information and de-identification

The footnote in the Australian National Statement on Research in Human Ethics, which excludes the use of certain terms for defining data, highlights how problematic the terminology surrounding de-identification is, and serves to create a complex and moving target:

> The National Statement does not use the terms "identifiable", "potentially identifiable", "re-identifiable", "non-identifiable" or "de-identified" as descriptive categories for data or information due to ambiguities in their meanings. (National Health and Medical Research Council 2018)

Whilst the National Statement chooses not to use such terms regarding identifiability of data, some of those terms currently exist in the Privacy Act and are crucial to its application. The Privacy Act, in its definitions section, defines "de-identified" as "personal information [...] no longer about an identifiable individual or an individual who is reasonably identifiable". The definition in the Privacy Act of personal information is that it "means information or an opinion about an identified individual, or an individual who is reasonably identifiable". The attraction of de-identification of data is that it is then argued to no longer be personal information, and therefore not subject to any protections offered by the Privacy Act.

As a result of this avoidance of protection of "personal information" via "de-identification", there is a significant incentive for organisations to assert that their data is de-identified, since it frees them to use for it for secondary purposes without consent, and to on sell the data locally or internationally. As such, the definition of de-identified data goes to the very heart of what

[1] See https://panopticlick.eff.org/.

data is personal and therefore afforded protection under the Privacy Act.

Industry, government, and academia have enthusiastically embraced the term de-identification. Whether this enthusiasm stems from a lack of technical understanding, or because it makes more data freely available for commercial use and research, remains unclear. The term 'de-identified' appears in numerous privacy policies, legislative acts, and policy proposals, both globally and in Australia. Commercial organisations use the term to justify the sharing and keeping of data.[2] Recent legislation and government policy proposals have put de-identification at the heart of future data sharing initiatives.[3] In academia the term has been used to expedite ethics waivers and to facilitate to the collection of data without the consent of the data subject.[4]

There is awareness of this problem. In July 2019, the ACCC suggested updating the definition of personal information, noting the need to "[u]pdate the definition of "personal information" in the Privacy Act to clarify that it captures technical data such as IP addresses, device identifiers, location data, and any other online identifiers that may be used to identify an individual" (ACCC 2019 Digital Platforms Report). The question as to why de-identification remains so popular in spite of this awareness remains an open one. An optimistic view of the flourishing of 'de-identified; data would be that its rise in popularity is due to a gap between the understanding of the technologies and the capabilities of the technologies. A slightly more sceptical view is that the weaknesses are well understood, but the terminology of 'de-identification' provides access to data that if classified correctly, would require a) consent that would not be forthcoming, or b) complete lack of access to data due to privacy protections.

2.2. RE-IDENTIFICATION

Re-identification is the inverse of de-identification, it seeks to partially, or fully, identify records that belong to a particular individual. It does not necessarily require that the individual's name be recovered, although that is often the case (Culnane et al. 2019). Re-identification merely requires that the individual becomes "reasonably identifiable", as defined in the Privacy Act. This is an important distinction, since harm can occur a long time before an individual's name is recovered. For example, if it is possible to link two supposedly de-identified records to each other, it might be possible to gain additional insight about that individual without their consent. As already discussed, a good example of this is in targeted online advertising, using data collected from online activity. Most online advertising platforms do not seek to tie analytics data to a specific name, they merely need to tie it to a 'profile', unidentified specifically by name or address. That profile itself can have a randomly generated identifier that does not identify an individual explicitly. Provided the platform can determine unique characteristics between interactions, the profile can be built about an individual without ever holding traditional identifiers like name and address. By way of an example, the Terms and Conditions for Google Analytics expressly forbid the submission of Personally Identifiable Information (PII) to the platform.[5] If that is the case, how do the likes of Google create targeted advertising? In effect, if a subset of an individual's data provides a unique signature, it can be used to index that individual across multiple datasets and collections without ever knowing their name or other traditional identifier. This is the technological piece that industry, policy-makers and regulators either do not understand or are wilfully blind to in continuing to speak of "de-identified data".

The risks of re-identification are not new. Not only have there been numerous examples of its failure in practice (Narayanan and Shmatikov 2007; Hern 2014), it has also been evaluated in academic literature. Paul Ohm highlighted the problem in 2010 in his paper 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization', in which he states:

> *[...] we have made a mistake, labored beneath a fundamental misunderstanding, which has assured us much less privacy than we have assumed. This mistake pervades nearly every information privacy law, regulation, and debate, yet regulators and legal scholars have paid it scant attention. (Ohm 2010)*

Unfortunately, after nearly 10 years, we are still labouring beneath the same fundamental misunderstanding. There have been many examples of real-world privacy failures of data releases that have claimed to be de-identified. De-identification as term is known to be ambiguous, and there have been calls for a cessation in its use (National Health and Medical Research

[2] See: Telstra. 2019. Privacy Statement. https://www.telstra.com.au/privacy/privacy-statement. Accessed 12/12/19; Tinder. 2018. Privacy Policy. https://www.gotinder.com/privacy. Accessed 12/12/19; NAB. 2019. Privacy Policy. https://www.nab.com.au/common/privacy-policy. Accessed 12/12/19.

[3] See Treasury Laws Amendment (Consumer Data Right) Bill 2019 (Cth); Office of the National Data Commissioner. 2019. Data Sharing and Release Legislative Reforms Discussion Paper. https://www.datacommissioner.gov.au/resources/discussion-paper. Accessed 12/12/19.

[4] See Department of General Practice, University of Melbourne. 2018. Data for Decisions: Data Sharing Agreement Summary. https://medicine.unimelb.edu.au/__data/assets/pdf_file/0003/2733267/Summary-of-Agreement-for-Provision-of-Data.pdf. Accessed 12/12/19.

[5] See Google. 2019. Google Analytics Terms of Service. https://marketingplatform.google.com/about/analytics/terms/us/. Accessed 12/12/19.

ISSN: 1839-4183

Council 2018). However, in 2019 it is still being introduced as a privacy protection technique in Australian legislation without addressing this recognised ambiguity.[6]

2.3. OPERATIONALISING DE-IDENTIFICATION

The fundamental techniques used for de-identification have not changed much since Ohm wrote his paper in 2010. Some techniques are extremely cursory, for example removing overt identifiers like name and ID numbers. Others attempt to preserve the underlying data by hashing or encrypting such identifiers, as was seen in the case of the MBS/PBS release (Department of Health. 2016). More sophisticated techniques attempt to ensure data about individuals does not exhibit unique characteristics (Samarati and Sweeney 1998). Different approaches offer varying levels of protection. Approaches that have removed identifiers or have been encrypted have been shown to be particularly vulnerable to re-identification (Culnane, Rubinstein and Teague 2017; Hern 2014).

The lack of robust guidelines and advice has allowed the ambiguity around de-identification to grow. Even when official advice is given, it often remains at best, subjective, and at worst, contradictory. In the US context the *Health Insurance Portability and Accountability Act of 1996* (HIPAA) is often used as a baseline to define de-identification. It states

> *There are two ways to de-identify information; either: (1) a formal determination by a qualified statistician; or (2) the removal of specified identifiers of the individual and of the individual's relatives, household members, and employers is required, and is adequate only if the covered entity has no actual knowledge that the remaining information could be used to identify the individual. (Privacy Rule of Health Insurance Portability and Accountability Act of 1996, Pub L 104–191, 110 Stat 1936)*

The problem with approach (1) is that it does not define how the qualified statistician is supposed to make the determination that information has successfully been de-identified. Without robust methods having been established it seems unlikely that a qualified statistician would be in a position to make such a determination. Approach (2) has problems as well, in that it defines what is commonly referred to as safe-harbour de-identification. If the entity removes the specified attributes, they can claim the data is de-identified. There are 18 specified attributes, covering a wide range of identifiers like name and address, through to IP address and bio-metrics. However, when looking at the last listed attribute it is a catch-all attribute, that states "(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section" (HIPAA 1996). If applied faithfully, and of some significance is the fact that individuals would need to be indistinguishable from at least one other person in the dataset to meet the safe harbour requirement. In practice, this is rarely the case.

For example, when a dataset was released under the HIPAA rules for the Heritage Health Prize the de-identification process made assumptions that would seem to contradict (R), in that they assumed, "[a]n adversary would have background information about only a subset of the claims of a patient in the dataset" (El Emam 2012). It could be argued that the sequence of claims from a patient was a unique characteristic, and HIPAA is written in terms of uniqueness, not the feasibility of a re-identification attack, and therefore applying bounds on the attacker seems inconsistent with HIPAA. This problem was highlighted by Narayanan when evaluating the privacy protections of the release (Narayanan, A. 2011), yet the release is considered to have met HIPAA requirements.

The situation is no better in the Australian context, the Office of the Australian Information Commissioner (OAIC) released guidelines on performing de-identification through the De-Identification Decision Making Framework (DDMF) (OAIC 2017a). However, a number of definitions within that framework are problematic, in particular the attempts to define de-identification. At various parts in the framework de-identification is described in three different ways. The following are all taken from the framework:

1. "A common error when thinking about de-identification is to focus on a fixed end state of the data."
2. "De-identification, then, is a process of risk management but it is also a decision-making process: should we release this data or not and if so in what form?"
3. "De-identification is a process to produce safe data, but it only makes sense if what you are producing is safe useful data…"

Taken together, these three definitions indicate it is both an end-state – in form of safe data according to (3), and not an end state according to (1), and also a decision-making process and risk management process according to (2). Given such a contradiction in the guidelines it is little wonder that the application of de-identification in the Australian context has been so problematic (Culnane, Rubinstein and Teague 2019; Culnane, Rubinstein and Teague 2017).

[6] See Treasury Laws Amendment (Consumer Data Right) Bill 2019 (Cth).

2.4. DE-IDENTIFICATION CONCEPTUALLY

On a more conceptual level, most common privacy protection techniques aim to make an individual indistinguishable from a crowd. The idea is that if an individual is indistinguishable from a crowd then their individual privacy cannot be breached. Risks remain for the privacy of a group in its entirety, for example, common attributes across the entire crowd. Techniques have been proposed to address such risks, but they are relatively rarely deployed in practice (Machanavajjhala et al. 2007; Li, Li and Venkatasubramanian 2007). As such, when evaluating the effectiveness of de-identification, the primary concern is whether an individual is distinguishable from others in the dataset. If they are, then there is an argument that the individual is re-identifiable via the distinguishing characteristics. Whether that individual is reasonably identifiable – the threshold required in the Privacy Act – is a subjective decision based on the likelihood of a third-party having access to auxiliary data that contains those distinguishing characteristics as well. It is also often defined by those with a direct conflict of interest in accessing the data.

It should be immediately clear that cursory de-identification techniques that just removed traditional identifiers are extremely unlikely to create a de-identified dataset. They often exhibit high degrees of distinguishability due to nothing more than the underlying data itself. The data points that represent that individuals' actions, devices, location, etc. are often as effective, if not more effective, at identifying an individual as traditional identifiers are at identifying an individual. This stands to reason, since most, if not all, individuals do not have a physical twin replicating their activity in real-life, from purchases through to travel patterns, and medical history. As such, there can be no expectation that a digital twin will naturally materialise to provide a crowd in which to hide that individual's data.

More advanced techniques that aim to ensure an individual is indistinguishable, most commonly delivered via *k*-anonymity (Samarati and Sweeney 1998), should conceptually work. The very design of the technique is intended to produce a group of individuals that are indistinguishable from each other. If that is the case, why do applications of such approaches continue to fail?

3. MISCONCEPTIONS OF K-ANONYMITY

As discussed in Section 2.3, vulnerabilities around the privacy of groups within *k*-anonymity have been identified and further advances to the techniques suggested. We will focus on the issue of individual identifiability since if the techniques are not protecting individual privacy, they will not protect group privacy. Although we do not have a protection of group privacy in Australia, some countries, such as Germany, provide protections explicitly in their Constitution, understandable given their history (Kommers and Miller 2012).

3.1. WHAT IS K-ANONYMITY?

One of the advantages of *k*-anonymity is that its definition is remarkably simple and intuitive, "The anonymity constraint requires released information to indistinctly relate to at least a given number k of individuals" (Samarati and Sweeney 1998) with one essential assumption that they "...consider the data holder's table to be a private table PT where each tuple refers to a different entity (individual, organization, and so on.)"7 (Samarati and Sweeney 1998). The constraint effectively means that the data for each individual is contained in a single row of a table.

In essence, *k*-anonymity makes any individual within a dataset indistinguishable from *k*-1 other individuals in that dataset. By setting a suitable value of k it is possible to adjust the level of risk associated with the release. For example, if k was set to 7, even if an adversary knew all the identifying characteristics of an individual in the dataset, there would 7 matching individuals in the dataset and as such they would have at best a 1 in k chance of picking the correct individual.

Applying *k*-anonymity requires generalising or suppressing data until the dataset meets the *k*-anonymity. Generalisation involves groups similar values together into broader classes, for example, combining postcodes or states. Suppression involves removing all or part of the data, for example, changing date of birth from day, month and year to just year. We will not expand on those processes here, as they are not the primary area of concern.

3.2. QUASI-IDENTIFIERS

One area we consider to be a weakness in *k*-anonymity, which also introduces subjectivity, is the notion of

[7] Where PT is a Private Table, which can be thought of as the privately held dataset. A tuple is a sequence of values, for example, the data contained in a single row of a table could be considered a tuple.

ISSN: 1839-4183

quasi-identifiers. Quasi-identifiers are defined as attributes the data holder believes might be used for linking to external knowledge. In Samarati and Sweeney (1998), the authors make the assumption that it is possible for the data holder to accurately discern the quasi-identifier and by extension, the current and future state of the external data available to prospective attackers. Whilst that assumption may have been valid in 1998 when the paper was published, it is questionable whether it remains true today. The scale and scope of data that is collected, and more importantly held in private databases, vastly exceeds anything that was anticipated 20 years ago. As such, it would appear to be impractical for anyone to be able to accurately determine the state of external knowledge, and such considerations should be removed from the definition. We therefore propose a simpler definition of a quasi-identifier, one that is not dependent on determining the state of external knowledge, namely, an attribute is a quasi-identifier if it is not a randomly generated identifier8, and there exist any two individuals within the dataset for which that attribute is distinguishing. Crucially, it does not require the attribute to uniquely identify an individual across the entire dataset, only between two records within the dataset. In effect, unless the attribute is a constant across the entire dataset, it should be considered a quasi-identifier, with the exception of any randomly generated identifiers. The definition assumes a worst-case scenario in which other fields or methods have been used to reduce the anonymity set to just two individuals. Were that to have occurred, does there exist another individual within the dataset for which the attribute in question is distinguishing? If so, the attribute should be considered a quasi-identifier. This greatly simplifies the application and negates the need to perform cross-field analysis. It will result in almost all fields being considered quasi-identifiers. As such, this will provide stronger privacy protection, whilst also removing the subjectivity of determining quasi-identifiers.

TABLE 1. Example Dataset.

Name	Age	Gender	State	Illness
Alan	42	Male	NSW	Heart
Carol	36	Female	Vic	Viral-infection
Kate	30	Female	Tas	Viral-infection
Bob	50	Male	NSW	Heart
Pete	40	Male	WA	Heart
Alice	39	Female	Vic	Viral-infection
Steven	62	Male	NT	TB
Jack	61	Male	SA	TB
Kelly	42	Female	Vic	Lung
Catherine	48	Female	Tas	Lung

3.3. CROSS-SECTIONAL DATA

k-anonymity was proposed during a time when most data releases were cross-sectional. A cross-sectional release is one that represents a single point in time. A good example would be the census, at least prior to 2016 in Australia. Such data was intended to represent that single moment, i.e. census night, and was not representing an ongoing period of time, nor would it be added to at a later date. For example, one census would not be directly linked

TABLE 2. After k-anonymity has been applied (k=2).

Name	Age	Gender	State	Illness
527	>40 <=50	Male	NSW/WA	Heart
960	>30 <=40	Female	Vic/Tas	Viral-infection
675	>30 <=40	Female	Vic/Tas	Viral-infection
808	>40 <=50	Male	NSW/WA	Heart
937	>40 <=50	Male	NSW/WA	Heart
749	>30 <=40	Female	Vic/Tas	Viral-infection
175	>60 <=70	Male	NT/SA	TB
647	>60 <=70	Male	NT/SA	TB
221	>40 <=50	Female	Vic/Tas	Lung
628	>40 <=50	Female	Vic/Tas	Lung

[8] Random identifiers are excluded because they will be unique across the dataset, but do not assist in identifying the individual themselves. It is essential the identifiers are random and in no way deterministic, for example, they must not be a cryptographic hash based on any or all of the data in the record. Such identifiers must be generated by the releasing party and not a reuse of existing identifiers, for example, randomly assigned ID numbers that exist outside of the release.

to a subsequent census. In such datasets the constraint that each individual is represented by a single tuple, or row in the table, is often naturally occurring.

Table 1 shows a simple example dataset prior to having *k*-anonymity applied. We can see that each individual is represented by a single row. We can therefore apply *k*-anonymity on those rows to achieve the target k value, the result of which is shown in Table 2. For reasons of brevity we have selected k=2, normally a higher value of k would be selected. Explicit identifiers are replaced with random values.

As can be seen in Table 2 there is at least *k*-1 other individuals that share the quasi-identifiers, which we have taken to be Age, Gender, State and Illness.

3.4 LONGITUDINAL DATA

In contrast to cross-sectional data, a lot, if not the majority, of the data collected today is longitudinal. This is data about an individual over a period of time and is often collected on an ongoing basis. For example, transactions, locations, past purchases, and medical history, are all examples of longitudinal data. Such data very rarely naturally meets the *k*-anonymity constraint of having all data related to an individual in a single tuple or row. In fact, the natural way of storing such data is the exact opposite of this, each transaction or event is stored as a single tuple or row itself. As such, each individual may have many rows of data that belongs to them, these can number in the tens if not hundreds of thousands for fine grained datasets (Department of Health, 2016). Furthermore, different sources of data, already granular and frequent, are increasingly being matched with other sources of data, increasing the likelihood of re-identification.

It is therefore crucial that when applying *k*-anonymity the dataset is pre-processed so that it meets the constraint of having all data about an individual in a single row or tuple. This requirement is well-known, having been an assumption in Samarati and Sweeney (1998), as well as having been discussed in Kifer and Machanavajjhala (2011), and more recently in Torra (2017). Despite this, as we shall see in Section 3.5, it is still not being explicitly stated in some guidelines and academic papers. When this is not performed, any *k*-anonymity calculation is incorrect, since it ignores the context of individual events, treating them as independent, when they are not. In the worst case, it might even result in a single individual being represented by all k rows, in effect, they are their own crowd, which obviously offers no privacy protection.

TABLE 3. Example longitudinal dataset.

Name	Age	Gender	State	Illness
527	>40 <=50	Male	NSW/WA	Heart
960	>30 <=40	Female	Vic/Tas	Viral-infection
675	>30 <=40	Female	Vic/Tas	Viral-infection
808	>40 <=50	Male	NSW/WA	Heart
937	>40 <=50	Male	NSW/WA	Heart
749	>30 <=40	Female	Vic/Tas	Viral-infection
175	>60 <=70	Male	NT/SA	TB
647	>60 <=70	Male	NT/SA	TB
221	>40 <=50	Female	Vic/Tas	Lung
628	>40 <=50	Female	Vic/Tas	Lung
749	>30 <=40	Female	Vic/Tas	Viral-infection
688	>30 <=40	Female	Vic/Tas	Lung
527	>40 <=50	Male	NSW/WA	Lung
348	>40 <=50	Male	NSW/WA	Lung
937	>40 <=50	Male	NSW/WA	Lung
175	>60 <=70	Male	NT/SA	TB
647	>60 <=70	Male	NT/SA	TB
960	>30 <=40	Female	Vic/Tas	Lung

To demonstrate this, we return to our previous example from Table 2, but include some additional events to demonstrate a longitudinal dataset, as shown in Table 3. For example, Person 527 now has two events, a Heart event and a Lung event. If we treat the rows as independent, and we re-examine the *k*-value of Table 3 we appear to achieve the specified k value of 2.

However, such a result is incorrect, because we have not met the constraint of having all data about an individual in a single row or tuple. As a result, it becomes possible to break the apparent *k*-anonymity by doing nothing more than reorganising the data from row-wise

ISSN: 1839-4183

TABLE 4. Longitudinal Uniqueness.

Name	Age	Gender	State	Illness 1	Illness 2
527	>40 <=50	Male	NSW/WA	Heart	Lung
960	>30 <=40	Female	Vic/Tas	Viral-infection	Lung
675	>30 <=40	Female	Vic/Tas	Viral-infection	
808	>40 <=50	Male	NSW/WA	Heart	
937	>40 <=50	Male	NSW/WA	Heart	Lung
749	>30 <=40	Female	Vic/Tas	Viral-infection	Viral-infection
175	>60 <=70	Male	NT/SA	TB	TB
647	>60 <=70	Male	NT/SA	TB	TB
221	>40 <=50	Female	Vic/Tas	Lung	
628	>40 <=50	Female	Vic/Tas	Lung	
688	>30 <=40	Female	Vic/Tas	Lung	
348	>40 <=50	Male	NSW/WA	Lung	

to column-wise, or more precisely, to arrange the data as is required by the correct definition of *k*-anonymity (Samarati and Sweeney 1998). Table 4 shows the results of that reorganisation, and it becomes immediately obvious that multiple individuals are now distinguishable, despite the previous appearance to the contrary.

In this example, there were only two events. As the number of events rises, it becomes increasingly difficult (and some would argue virtually impossible) to deliver the necessary *k*-anonymity value (Aggarwal 2005). More attributes need to be generalised or suppressed, often to a greater extent, in order to be able to deliver the required indistinguishability of individuals. This can significantly reduce the utility of the data, potentially making the release worthless. Further challenges occur when the number of events each individual has varies, since the mere count of events can become a distinguishing factor.

It is important to note that the failure is not with the original definition of *k*-anonymity, it is the incorrect application of it. If applied correctly it would rapidly become evident that large-scale longitudinal release, like the MBS/PBS release9, cannot be made safe without destroying its utility. This raises the question as to why the technique is being applied incorrectly?

3.5. BAD ADVICE AND INCORRECT DEFINITIONS

The incorrect application of *k*-anonymity is not new, however, in the Australian context it is particularly problematic, because official guidelines, in the form of the DDMF, perpetuate the incorrect application. Firstly, the definition they provide of *k*-anonymity, in the appendices, is incorrect

> *A dataset is regarded as k-anonymised if – on all sets of key variables – each combination of possible values of those variables has at least k records that have that combination of values. In essence, this gives a standard for data to be considered safe (OAIC 2017b).*

The definition makes no mention of individuals, instead referring to records. It also makes no mention of the requirement that all the information related to an individual reside in a single tuple or row. The usage of the term "record" is also problematic, since the term typically refers to a single row in a database, and as such, longitudinal data will consist of many records for a single individual. Someone applying *k*-anonymity on longitudinal

[9] Department of Health. 2016. Public Release of Linkable 10% sample of Medicare Benefits Scheme (Medicare) and Pharmaceutical Benefits Scheme (PBS) Data. http://www.pbs.gov.au/info/news/2016/08/public-release-of-linkable-10-percent-mbs-and-pbs-data. Accessed 12/12/19.

dataset according to the definition in the de-identification decision making framework will likely end up with an incorrect result that will deliver considerably less privacy than appears.

This is a common definitional error found in academic publications (see El Amam and Dankar 2008). However, such papers apply the technique to cross-sectional data, in which a single record naturally relates to an individual. As such, it is possible to see how a slight change in terminology, which made no material difference in cross-sectional data, came about, and how the implications of that change were not fully appreciated when applying the technique to longitudinal data. The fact the mistake occurred regarding the definition of identifiable data is perhaps understandable, the fact it persists in the face of repeated real-world failures is not.

In addition, the volume and speed at which data is being collected makes this not only an issue in the present, but also requires further contemplation as volumes of data are on the rapid rise, both in speed and quantity. Data is increasingly linked in ways that show patterns and identify individuals and groups that were not previously possible. The protection of privacy in the face of future developments also needs to be regulated and upheld in policy moving forward.

3.6 K-ANONYMITY APPLICATION

The primary issue discussed in this paper is one of incorrect definitions and misconceptions of the application of *k*-anonymity to longitudinal data. Even with the clarified definitions, the fundamental challenge of effectively performing *k*-anonymity remains. The challenges of determining a suitable *k*-value, for example, is not resolved. Achieving the appropriate balance between privacy and utility is well-established challenge. There are two issues discussed in this paper. The first is the lack of clarity around the definitions of privacy, and that these definitions are not reflected in technical practices. We raise the point that subjectivity and lack of understanding of technical approaches should not be used to hide behind when neglecting to uphold fundamental privacy principles.

The second issue is the challenge posed by determining a suitable *k*-value, or the challenge of anonymisation itself. Even when a suitable *k*-value has been determined, optimally applying suppression or generalisation to achieve it remains a computationally hard problem (Meyerson and Williams 2004; Aggarwal et al., 2005), and one that we do not address in this paper. It is important to note that even when correctly applying *k*-anonymity, there is no guarantee that an acceptable release can be generated to comply with existing privacy laws. Appliers of *k*-anonymity should accept that such an outcome is possible and realise that some data sets may not be suitable for release using *k*-anonymity.

4. CURRENT AND FUTURE PROBLEMS

4.1. CURRENT STATE OF PLAY IN AUSTRALIA

In 2008, the Australian Law Reform Commission (ALRC) was tasked with considering whether a statutory cause of action for serious invasion of privacy would be beneficial to the Australian community (ALRC 2014). In its report, *For Your Information: Privacy Law and Practice*, the ALRC focused on data protection: information collection, access and use. The ALRC found in the affirmative and noted that, as a recognised human right, privacy was paramount. Cost and convenience should be subservient to privacy, and although there would often be competing rights, and everyone recognises that privacy is not an absolute right, it must always be thoughtfully and carefully balanced against those other competing rights, and not automatically subsumed by them. Privacy should be given greater attention and protection, according to the ALRC report.

In 2017, only 9 years later, but a world away in terms of the media and digital landscape, a further inquiry was launched by the Australian Competition and Consumer Commission (ACCC). The ACCC inquiry was tasked with looking at the effects that digital search engines, social media platforms and other digital content aggregation platforms have on competition in media and advertising services markets. In particular, the ACCC inquiry looked at the impact of digital platforms on the supply of news and journalistic content and the implications of this for media content creators, advertisers and consumers. The ACCC Inquiry, like the ALRC, recommended the introduction of

ISSN: 1839-4183

a statutory tort for serious invasions of privacy. The ACCC also recommended that individuals be given rights to sue for breaches of the Privacy Act (which would stand separately to the statutory tort of privacy also recommended). Each of these remedies would be available for different wrongs, under different circumstances.

Most recently, in July 2019, the final recommendations of the ACCC report were handed down. Recommendation 16 suggests updating the definition of personal information, noting the need to "[u]pdate the definition of "personal information" in the Privacy Act to clarify that it captures technical data such as IP addresses, device identifiers, location data, and any other online identifiers that may be used to identify an individual." The most recent 2019 ACCC Digital Platforms Inquiry mentions privacy over 1000 times (ACCC 2019).

4.2. FUTURE PROBLEMS

The ongoing technically incorrect definition in the DDMF is a significant problem as it creates the impression that unit record level longitudinal data about individuals can be de-identified, when in practice this is not the case. This misconception is continuing to perpetuate in new policy and legislation. The Consumer Data Right legislation and Rules both include references to de-identification[10]. In the case of the rules for banking data, this would be de-identification of up to a year's worth of transaction data. The suggested method for performing such de-identification is the DDMF. As such, we can be fairly certain it will not provide the necessary privacy protection.

Looking further ahead, the governments flagship data sharing and release policies also make reference to de-identification and the DDMF as a recommended approach. This has created a situation in which the Australian government is making policy and legislation on the basis of terminology that reflects a technological fiction. Ohm's words have never been truer in an Australian context: misconceptions about de-identification continue to pervade nearly every information privacy law, regulation, and debate. Until this misconception is corrected, the privacy of individuals will continue to be compromised.

4.3. POLICY ADVICE AND NEXT STEPS

The first point we make is that regulators and policy-makers need to contemplate the privacy of individuals as well as the privacy of groups. This article examines how incomplete and ambiguous definitions, as well as poor applications, can lead to data being asserted as de-identified, when in fact it remains, "personal information" which is information that is, "about an identifiable individual or an individual who is reasonably identifiable." It thus follows that allegedly "de-identified data" may still need to comply with the protections of the Privacy Act. Regulators and policy-makers need to approach all data that is categorised as "de-identified" with scepticism and assume that the Privacy Act still applies unless unequivocally demonstrated it does not.

[10] ACCC. 2019. CDR Rules (banking). https://www.accc.gov.au/focus-areas/consumer-data-right-cdr-0/cdr-rules-banking. Accessed 12/12/19.

The second point is that regulators and policy-makers need to address the increasing technical capacity to link sources of otherwise benign information, which may provide a profile of sensitive and personal (and most importantly, identifiable information). In light of these advances, what does a right to privacy even mean? What is the standard for privacy when so much data is collected that can be linked and traced back to an individual? Regulators and judges alike seem reluctant to wade into this increasingly complex field, and yet theirs will be the role to establish the limits and protections that protect citizens into the future.

What does seem clear is the rising crescendo of experts calling for a greater protection of privacy – whether by statute or by tort. This increase in demand seems matched only by the lack of appetite by regulators and judges to respond at a time when the fundamental protections afforded by the Privacy Act are undermined and ignored, either wilfully or negligently. Either way, policy-makers, regulators and judges urgently need to grapple with the technologies and to respond with meaningful law and policy that is clear and protects, not undermines, existing rights.

5. REFERENCES

1. ACCC. 2019. *CDR Rules (banking)*. https://www.accc.gov.au/focus-areas/consumer-data-right-cdr-0/cdr-rules-banking. Accessed 12/12/2019.
2. ACCC. 2019. Digital Platforms Inquiry: Final Report. https://www.accc.gov.au/system/files/Digital%20platforms%20inquiry%20-%20final%20report.pdf. Accessed 12/12/2019.
3. Aggarwal, C.C., 2005. "On k-anonymity and the curse of dimensionality." VLDB '05: Proceedings of the 31st International Conference on Very Large Data Bases, August, VLDB, pp. 901–909.
4. Aggarwal, G. ,Feder, T., Kenthapadi, K. , Motwani, R,, Panigrahy, R. and Thomas, D. and Zhu, A. (2005). "Approximation Algorithms for k-Anonymity". Proceedings of the International Conference on Database Theory, ICDT 2005, January 5-7, Edinburgh, UK.
5. Australian Law Reform Commission. 2014. Serious Invasions of Privacy in The Digital Era (ALRC Report 123). https://www.alrc.gov.au/publications/1-executive-summary/should-new-tort-be-enacted. Accessed 12/12/2019.
6. Barbaro, M. and T. Zeller Jr. 2006. "A Face Is Exposed for AOL Searcher No. 4417749." New York Times, New York, 9 August. https://www.nytimes.com/2006/08/09/technology/09aol.html. Accessed 12/12/2019.
7. Culnane, C., B. I. P. Rubinstein and V. Teague. 2017. "Health Data in an Open World." arXiv https://arxiv.org/abs/1712.05627. Accessed 12/12/2019.
8. Culnane, C., B. I. P. Rubinstein and V. Teague, 2019. "Stop the Open Data Bus, We Want to Get Off." arXiv https://arxiv.org/abs/1908.05004. Accessed 12/12/2019.
9. Department of General Practice, University of Melbourne. 2018. Data for Decisions: Data Sharing Agreement Summary. https://medicine.unimelb.edu.au/__data/assets/pdf_file/0003/2733267/Summary-of-Agreement-for-Provision-of-Data.pdf. Accessed 12/12/2019.
10. Department of Health. 2016. Public Release of Linkable 10% sample of Medicare Benefits Scheme (Medicare) and Pharmaceutical Benefits Scheme (PBS) Data. http://www.pbs.gov.au/info/news/2016/08/public-release-of-linkable-10-percent-mbs-and-pbs-data. Accessed 12/12/2019.
11. El Emam, K. and F. K. Dankar. 2008. "Protecting Privacy Using k-Anonymity." Journal of the American Medical Informatics Association, 15 (5): 627–637.
12. El Emam, K., et al. 2012. "De-identification methods for open health data: the case of the Heritage Health Prize claims dataset." Journal of Medical Internet Research, 14 (1): e33.
13. Google. 2019. Google Analytics Terms of Service. https://marketingplatform.google.com/about/analytics/terms/us/. Accessed 12/12/2019.
14. Health Insurance Portability and Accountability Act of 1996, Pub L 104–191, 110 Stat 1936.
15. Hern, A. 2014. "New York taxi details can be extracted from anonymised data, researchers say." The Guardian, online, 28 June. https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn. Accessed 12/12/2019.
16. Kifer, D. and Machanavajjhala, A. 2011. "No free lunch in data privacy." Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, Athens, Jun 12-16, ACM, pp. 193-204.
17. Donald Kommers, D. and Russell R. Miller. 2012. The Constitutional Jurisprudence of the Federal Republic of Germany. Durham: Duke University Press.
18. Li, N., T. Li and S. Venkatasubramanian. 2007. "t-closeness: Privacy beyond k-anonymity and l-diversity." 2007 IEEE 23rd International Conference on Data Engineering. Istanbul, IEEE, pp. 106-115.

ISSN: 1839-4183

19. Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M., 2007. "l-diversity: Privacy beyond k-anonymity". ACM Transactions on Knowledge Discovery from Data, 1 (1): 3-es.
20. Meyerson A, Williams R, 2004. "On the complexity of optimal k-anonymity." Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 223-228.
21. Moller, R. 2019. Vox Media Privacy Policy Explained: What We Know about You. https://www.vox.com/recode/2019/12/10/20962868/vox-media-privacy-policy-explained-what-we-know-about-you. Accessed 12/12/2019.
22. NAB. 2019. Privacy Policy. https://www.nab.com.au/common/privacy-policy. Accessed 12/12/2019.
23. Narayanan, A. 2011. "An adversarial analysis of the reidentifiability of the heritage health prize dataset." Unpublished manuscript.
24. Narayanan, A. and V. Shmatikov. 2007. "How To Break Anonymity of the Netflix Prize Dataset." arXiv https://arxiv.org/abs/cs/0610105.
25. National Health and Medical Research Council. 2018. National Statement on Ethical Conduct in Human Research (2007) – Updated 2018). https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018. Accessed 12/12/2019.
26. Office of the Australian Information Commissioner. 2017a. De-identification Decision-Making Framework. https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-decision-making-framework/. Accessed 12/12/2019.
27. Office of the Australian Information Commissioner. 2017b. De-identification Decision-Making Framework Appendices. https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-decision-making-framework/. Accessed 12/12/2019
28. Office of the Australian Information Commissioner. 2018. De-identification and the Privacy Act. https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-and-the-privacy-act/. Accessed 12/12/2019.
29. Office of the National Data Commissioner. 2019. Data Sharing and Release Legislative Reforms Discussion Paper. https://www.datacommissioner.gov.au/resources/discussion-paper. Accessed 12/12/19.
30. Ohm, P. 2009. "Broken promises of privacy: Responding to the surprising failure of anonymization." UCLA Law Review, 57 (6): 1701-1778.
31. Privacy Act 1988 (Cth). No. 119. Compilation No. 82 https://www.legislation.gov.au/Details/C2020C00025 , Accessed 16/04/2020.
32. Samarati, P. and L. Sweeney. 1998. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression". Technical Report SRI-CSL-98-04. Computer Science Laboratory, SRI International.
33. Telstra. 2019. Privacy Statement. https://www.telstra.com.au/privacy/privacy-statement. Accessed 12/12/2019.
34. Tinder. 2018. Privacy Policy. https://www.gotinder.com/privacy. Accessed 12/12/2019.
35. Torra, Vicenç, 2017. Data privacy: foundations, new developments and the big data challenge. Cham: Springer International Publishing.
36. Treasury Laws Amendment (Consumer Data Right) Bills Digest No. 68, 2018–19 (Cth). https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/bd/bd1819a/19bd068 Accessed 16/4/2020.
37. Zuboff, S. 2019. The Age of Surveillance Capitalism. London: Profile Books.