Response to the Productivity Commission's draft report on data availability and use

Dr Chris Culnane	Dr Benjamin Rubinstein	Dr Vanessa Teague
cculnane@unimelb.edu.au **+61 3 8344 1408**	brubinstein@unimelb.edu.au +61 3 9035 6657	vjteague@unimelb.edu.au +61 3 8344 1274
Department of Computing and Information Systems, The University of Melbourne **Parkville, VIC 3010, Australia**		

This submission addresses de-identification, re-identification and privacy in data policy. We would be happy to discuss or expand upon these issues.

Dr Chris Culnane is a research fellow also specialising in cryptographic protocols for electronic voting. He was the lead developer on the Victorian Electoral Commission's vVote project, the first end-to-end verifiable voting system to run at a state level anywhere in the world.

Dr Ben Rubinstein is a senior lecturer at the University of Melbourne. His expertise is in machine learning, statistics and data privacy for example through differential privacy. He previously worked at Microsoft Research and a number of other Silicon Valley companies.

Dr Vanessa Teague is a senior lecturer at the University of Melbourne. Her expertise is in cryptographic protocols, particularly those for verifiable elections and other public processes. She is an Advisory Board member of Verified Voting, a non-partisan non-profit US organization that advocates for legislation and regulation that promotes accuracy, transparency and verifiability of elections.

Ours was the team that in September 2016 recovered supplier IDs in the 10% MBS/PBS longitudinal dataset, released by the federal Department of Health. Our results, responsibly disclosed to the relevant Department, immediately preceded the announcement by Government of the re-identification bill amending the *Privacy Act 1988* to criminalise re-identification.

Executive Summary

- De-identification doesn't work—it does not solve the problem of making sensitive unit-record-level data available openly but safely. The ease of re-identification is context and time-dependent – relying on de-identification for privacy is unwarranted because re-identification only gets easier as more auxiliary information is made more widely available for linking. Data privacy and utility are at odds in this context. If not enough information is removed, people's records may be easily re-identified; if too much is removed, scientists may find it useless for research. For some datasets, there may be no point that is acceptable by both criteria.

- The report's trust framework however is well-suited to mitigating privacy risks while maintaining acceptable levels of utility. Sensitive data could be shared with trusted users who have demonstrated that they can keep it secure.

- The ease of re-identification depends very much on what auxiliary information is available to an attacker. De-identification is not defined precisely in the report, the *Privacy Act 1988*, or its proposed amendment; it relies on the vague notion of whether a person can "reasonably" be identified.

- Throughout the draft report and many other government documents, "de-identified" is used to mean two very different things: that a *process* has been performed, and that a secure *state* (in which people cannot reasonably be identified) has been achieved. The two should not be confused: intending to make something secure does not actually make it secure.

- Harm from improperly de-identified datasets may not require unique re-identification. Sensitive information about an individual might be revealed without linking them to one record. For example, one person could be narrowed to two records that both have a certain trait.

- We support progress on compliance guidelines for de-identification but warn against over reliance. The guidelines should be public.

- We are concerned about unintended consequences of the proposal to amend the *Privacy Act 1988* to criminalise re-identification. Responsible re-identification is important for improving data privacy; we recommend that the bill instead criminalise the misuse and distribution of re-identified data.

- There are good reasons for a clear separation of roles between the OAIC (regulation, strategy and dispute resolution) and the NDC (data release process and policy). There are bound to be more privacy breaches: it is important that the authority investigating them is not also one of the parties that might have unintentionally contributed.

- Accreditation and privacy protection requires deeply technical skillsets: it is paramount that sensitive releases are properly examined. Proper resourcing should permit the NDC to act as a central hub, under guidance of a technical advisory board.

- The algorithms used for de-identification, encryption and other privacy protections should be made openly available for public scrutiny, preferably before they are used, so that errors and weaknesses can be found and fixed. This would be good for security and good for public trust.

Introduction and Summary

Data driven policy and research holds huge potential benefits for society. We wholeheartedly support the idea of data-driven policy and research. We welcome a number of the proposals outlined in the *Data Availability and Use* report, particularly the proposed trust framework for managing the release of data. However, de-identification followed by open publication is not a secure way to make sensitive unit-level records[1] available for research or policy.

There is clear evidence in the report of the hurdles, bottlenecks and time delays that researchers have faced when trying to access data. It is clearly not desirable to have researchers waiting years for access to data, particularly when such data has the potential to save lives. It is tempting to pursue an open data solution to the problem, in that making all data open solves the problems of access. As tempting as that may be, it would be a mistake to pursue such a path. It does not solve the original access problem, and it creates a new privacy problem.

De-identification works by trying to obscure or remove some part of unit-record data. At one extreme, all meaningful information could be removed; at the other, raw records could be published and easily re-identified. The question is whether there is a useful middle ground. Are there techniques which allow a reasonable protection of privacy while also providing good data for research? For some datasets, such as unit-level health, mobility or Centrelink records, this is probably not possible. Public policy should certainly not be based on the assumption that secure techniques exist.

The challenge is finding the balance between providing access to rich datasets for those whose access benefits the public, whilst protecting the privacy of individuals. We should not derive our direction from the status quo, and should instead set out a plan to improve privacy, and educate the public about its importance. The fact that many individuals overshare on social media, and are currently exchanging vast amounts of personal information for incidental gain, should be a cause for concern, not a reason for the government to publish even more of their data. Progress and good policy cannot be made without seriously and comprehensively addressing privacy concerns. The movement to boycott the census *or to provide inaccurate data* shows that people will provide truthful data only if they are confident it will be protected. Indeed, privacy problems can undermine productivity, for example because people do not feel free to seek the health care or psychiatric support they need.

We support the recommendation for using a trust framework to protect privacy, and argue that such an approach offers a solution to the problems of access to high quality data whilst still providing adequate privacy protections.

The term "de-identification" has been misconstrued and is now so ambiguous that use of the word itself should be avoided. A well-meaning official carefully following a de-identification *process* might fail to achieve the secure *state* in which individuals cannot be "reasonably identified". Terminology should be chosen very carefully: if data really cannot reasonably be used to identify individuals, then it might be safe to post it online. This should be called "non-identifiable" - it should not be assumed to have been achieved just because de-identification guidelines were followed. Something that was de-identified one day might be easily re-identifiable the next because more data has become available for linkage. As such, de-identification should not be relied upon to provide ongoing privacy protection. It might be useful in a context with a lesser attack model, for example when a dataset

[1] Unit-level records are individual records (about individual people, or other entities), rather than records made from aggregating other data using sums, averages, etc.

has been given to a trusted researcher who has a good privacy framework in place. This sort of partial de-identification is merely a tool for use as part of a larger process. This view is not incompatible with many of the recommendations within the report, and is entirely compatible with the proposed trust framework. However, it does impact on the classification of data and the determination of what data can be made safely available to the public. We shall expand on this in later sections.

The proposed amendments to the *Privacy Act 1988*, to criminalise re-identification, both stifle legitimate public interest research, and fail to adequately protect individuals. Privacy would be better protected by a law that focused on preventing harm caused by inadequately de-identified public data, while encouraging examination of privacy protections with responsible disclosure. As drafted it risks silencing the messenger, whilst failing to protect against malicious entities.

We attach an appendix with observations about the wider data sharing environment. We are largely in agreement with a number of the recommendations in the report around this topic, but again differ on some definitions, and in some cases would recommend stronger consumer protections to be created. We reiterate the ease with which complex records can often be re-identified, even if they have supposedly been de-identified, and especially by corporations with large databases of auxiliary information. We also point out that harm can be done without explicit re-identification, for instance by discriminating against someone by learning extra information about them but without necessarily learning their identity. The current situation creates an inequality between corporations and consumers. As corporations build bigger and better profiles of their customers the inequality grows. There is a very real danger that a situation will develop in which consumers will be routinely exploited due to the imbalance of information possessed by the respective parties. We would recommend broader definitions and additional protections to either redress the imbalance or restrict the inequality.

De-identification

In this section we discuss why de-identification does not provide high quality research data, before looking at the fallacy surrounding it, followed by a discussion on how its loose definition, and misappropriation, leads us to recommend avoiding its use altogether in the open data context.

De-identification – Impact on quality

De-identification methodologies involve a number of different techniques, including encrypting unique identifiers, removing highly identifiable fields and perturbing data to attempt to reduce uniqueness. Each of the measures undertaken will impact on the quality of unit-record data, and the types of research that can be conducted on it. This pits privacy against utility in a battle in which there are no winners. At the extremes, maximum utility is achieved by releasing all data with no adjustments, whilst maximum privacy is achieved by deleting all data and releasing nothing. Most dataset releases settle on a point somewhere between the two. Unfortunately, this often results in privacy not being adequately protected, whilst simultaneously failing to provide the necessary utility for high quality research.

Even those who advocate de-identification as a legitimate tool for privacy protection agree that when releasing data publicly the de-identification, and therefore the loss in utility, must be greater. Releasing essential research data in an open data context will not provide the optimal level of utility, but will still risk privacy.

The fallacy of de-identification

De-identification is often portrayed as methodology to transform identifiable information into a form in which it is no longer identifiable. De-identification is at best obfuscatory, not transformative, always leaving a risk of re-identification. The debate between de-identification advocates and opponents is not on the basis of whether re-identification is possible, but on the acceptable level of re-identification risk, and more broadly, whether it is even feasible to quantify that risk.

De-identification is fallible due to the dynamic nature of the data release environment. Everyday more data is added to the public domain, whether it be via new datasets, social media, news stories, or any other updates to publicly accessible data. What data will be released, and when, is neither predictable, nor controllable. A hypothetical calculation of re-identification risk can only be valid on the day it is performed, because it makes specific assumptions about the auxiliary information available to the attacker. The data release environment will be changing every day and the risk of re-identification will change with it.

Given that the risk of future re-identification cannot be accurately predicted, and is dependent on uncontrollable events, it is unadvisable to publicly release data whose privacy is dependent on the accuracy of that predicted risk. What is more, publicly released data cannot be recalled. Its release is a onetime event, once public it is public forever.

The notion of the fallacy of de-identification is not new, it is widely discussed by Ohm in his paper *Broken Promises of Privacy: Responding to the surprising failure of anonymization*. [1]

Defining de-identification

At the root of the problem is a lack of a strong definition of what de-identification is. The *Privacy Act 1988* states “personal information is de-identified if the information is no longer about an identifiable individual or an individual who is reasonably identifiable.” The use of 'reasonably' in the definition is problematic, firstly, how does one judge what "reasonably identifiable" even means? Secondly, in this context, the reasonable behaviour is being projected onto the adversary, *i.e.* an attacker will only try to re-identify the data using reasonable means. But what is a “reasonable” attacker on the Internet? Such attackers may be persistent, ingenious, well-funded and in possession of vast collections of auxiliary data.

One important point, which is often overlooked, is that the Privacy Act defines an outcome not a methodology. Despite this, the term de-identification has been misappropriated by vendors, and some researchers, to refer to a methodology to be performed on identifiable data to allow it to be publicly released. The *outcome* that people cannot be reasonably identified may not be successfully achieved by a suggested method.

This combination of loose and conflicting definitions creates ambiguity, allowing an entity to use a de-identification tool to comply with legislation, whilst not actually protecting the privacy of the data. This is a recognised problem, the Australian National Data Service (ANDS) website states:

> *“This National Statement avoids the term 'de-identified data', as its meaning is unclear. While it is sometimes used to refer to a record that cannot be linked to an individual ('non-identifiable'), it is also used to refer to a record in which identifying information has been removed but the means still exist to re-identify the individual. When the term 'de-identified data' is used, researchers and those reviewing research need to establish precisely which of these possible meanings is intended.*” [2]

The use of “de-identified” in reference to data that may be re-identifiable should end. ANDS suggests a preference for the term “non-identifiable”, which is defined in the *National Statement on Ethical Conduct in Human Research* as:

> *“Non-identifiable data, which have never been labelled with individual identifiers or from which identifiers have been permanently removed, and by means of which no specific individual can be identified.”* [3]

This is a much stronger, and in our opinion, superior definition. It is not subjective and is absolute in its requirement that no specific individual can be identified. We would recommend that use of the term "de-identified" be dropped and replaced with "non-identifiable" when referring to data that is being made open. We recognise that this will have significant impact on the public release of unit-record level datasets, however, we believe it would be compliant with the trust framework proposed in the report.

De-identification – does it still have a purpose?

De-identification can still serve a purpose—it is a useful technique for protecting against accidental or observational privacy breaches. For example, there will be datasets for which identity is not desired, nor sought, and for which de-identification is useful to allow easy access and analysis while mitigating fear of accidental identification. However, the privacy of the dataset should not be reliant on the de-identification. Instead, it should be based on the established trust framework. This is largely covered already by Figure 9.1 in the report. However, we would recommend a slight change in how the trust framework is described. Rather than classifying data as identifiable or de-identified, the data should be classified based on sensitivity.

Re-identification vs Breaches

The report suggests that re-identification is less common than data breaches, and as such, is not necessarily a primary concern. However, the two are not equivalent, and the lack of evidence of re-identification does not mean it is not occurring. Unlike many data breaches, where database access and exfiltration are logged by a system, re-identification occurs in private. There is no way of knowing if or when it is occurring. The only time it becomes apparent is when someone discloses to the public, or the releasing agency, that they are able to re-identify the data. As such, the pattern of re-identification only occurring as a result of researcher or journalistic endeavour is a misconception. Researchers and journalists are motivated by public interest and as such disclose their findings. An entity motivated by profit or malice is unlikely to disclose their actions. As such, it is highly likely the re-identification has gone on in other datasets, and could well currently be going on, we just do not know about it.

The current rate of sensitive data breaches should be a matter of serious concern. If anything this argument should be used to justify greater efforts at ordinary security of systems that hold sensitive data. It is not a reason to make even more sensitive data available in an even less secure way.

Harm without Full Re-identification

It is often assumed that for harm to occur an individual needs to be uniquely identified within the dataset. In reality the threshold for harm occurs considerably earlier than individual identification. Partial re-identification can occur when a class or category of individuals can be identified, and some characteristic can be assigned to that category. For example, it might be possible to narrow an individual down to a set of 100 records. That in itself does not yield personally identifiable information, but if you know that all 100 suffer from heart disease you are able to derive sensitive information without having to uniquely identify an individual. Even if you only know that the

instances of heart disease within that class are greater than the national average, it could adversely affect the individual.

As a result, data that remains technically de-identified can still cause harm to an individual. With regards to the definition of consumer data in the report (p. 345) we would argue that maintaining an exemption for data that is not strictly identifiable presents a risk that companies will classify their customers into favourable classes and avoid the restrictions that would be imposed on strictly identifiable data. We would recommend a broader definition to require companies to provide access to the data on which a decision related to the consumer has been made, whether or not that information is identifiable. This will protect against misclassification, which can be as equally damaging as actions taken against an individual.

De-identification Guidelines and Certification

We support the creation and publication of best practice guidelines on de-identification. However, we do not believe de-identification can ever be sufficient to truly protect privacy for unit-record data. As such, the guidance must make clear that de-identification must be combined with a suitable trust framework to protect privacy. It is also essential to emphasise the de-identification is not a static process, it requires constant re-evaluation and updating in order to keep up with the dynamic nature of the data release environment.

The notion of accreditation and certification is also troublesome in this context. The methodologies used should be open and published prior to any publication of data, to allow expert and community feedback on the proposed methods. Since de-identification is not a panacea and is extremely context dependent, it is difficult to see how a set of certification criteria can be constructed that adequately capture the task at hand. A poorly constructed set of certification criteria runs the risk of making matters worse, particularly if it permits technical compliance whilst not delivering adequate privacy protections. It also runs the risk of motivating data guardians to comply with a set of guidelines in order to achieve certification, when instead they should be motivated to protect the privacy of their data.

Privacy Act 1988

Proposed Changes to the Privacy Act

The recently proposed changes to the Privacy Act, regarding re-identification, are unfortunately not productive. They run the risk of stifling public interest research, whilst not protecting against actions that are detrimental to individuals. Legislation should consider intent and harm when restricting actions. By outlawing the act of re-identification of government datasets, responsible researchers are prevented from being able to analyse and report failures. We argue that such data should not be released publicly in the first place, but if it is, preventing researchers from finding and reporting failures is highly undesirable. The notion that researchers should apply for prior permission also fundamentally misunderstands re-identification research. Re-identification methods are not particularly complicated, they contain very little novel research. There is a misconception that the analysis is somehow new or enabled by recent advances in technology. The reality is often that it is something that could have been done since the dawn of large databases. As such, conducting re-identification analysis is unlikely to lead to publications or substantial research funding. Researchers undertake the analysis out of public interest, and largely outside of existing research commitments. Increasing the burden and costs of undertaking such actions will make them infeasible and the result will be researchers not providing that public service. This will have no impact on malicious entities undertaking re-identification, which will continue unabated. The result will be a reduction in

situational awareness and increasing risk, which will not serve the public interest. Our specific recommendations are in our submission to the Senate inquiry[2].

Data Release Process and Infrastructure

Trusted Users

There should be multiple categories of trusted users, with different levels of trust. Universities, government agencies, and state and territory authorities should be considered to be public interest research organisations (select other organisations may also be added to this list). Such organisations should have wide access and few restrictions on the analysis they can run on the data. However, strict limitations should be placed on how those organisations store, protect, and publish findings from the data.

Organisations should demonstrate they have suitable infrastructure to secure and protect any identifiable data they have. This could be via offline facilities. Funding should be provided to allow Universities to establish such facilities to permit efficient and safe access.

Corporations and not-for-profits should not be assumed to be operating in the public interest. As such, their access to datasets should be determined on a case by case basis, evaluated on the public interest of their proposals. There should be restrictions on use of the data and deletion of such data at the culmination of any project.

NDC, Responsibilities and Accreditation

We would recommend the NDC take on a larger role, providing best practice guidelines, data releases for government departments and accreditation of trusted users. As such, it will require a significant technical capability in order to deliver at high service levels. For example, the accreditation of trusted users will require an understanding of computer security, including encryption, cybersecurity, risk management, and access control. In addition its role in overseeing the release of public data will require capabilities in both generating and evaluating complex data analyses of prospective datasets, touching on record linkage, statistics, machine learning and information theory. Expertise in the relevant skills could be concentrated in once place.

The OAIC should maintain its distinct role in regulation, strategy and dispute resolution, expanding to cover the actions of the NDC. Requiring the OAIC to also engage in accreditation may weaken its dispute resolution role, particularly if there was an overlap in enforcement and accreditation. The issue is compounded by trying to produce accredited best practice. The field is not mature enough to have established best practice and creating the impression that it exists risks shifting responsibility for failure from the data guardian to the accreditation body. There are bound to be more privacy breaches – there needs to be a clear separation between the investigating authority and the parties who might have contributed. The OAIC must be shielded from the risk of shifting responsibility, allowing it to effectively enforce and protect the privacy of Australians.

Privacy cannot be achieved through compliance with checklists; it requires careful analysis of each prospective dataset and its position in the wider data environment. The oversight role will require awareness of advancements in both privacy protection and privacy incursion techniques, in order to keep best practice ahead of the game. We would recommend the creation of an independent

---

2 http://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Legal_and_Constitutional_Affairs/PrivacyRe-identification/Submissions.

advisory board in order to provide expertise from a range of different sectors (industry, academia, and advocacy).

Failure to correctly resource such an organisation runs the risk of creating a new, and worse, bottleneck on accessing data. It is also important to recognise that trusted users should be re-accredited at regular intervals to ensure the trust framework remains secure into the future.

It is essential that government departments are adequately resourced and supported by the NDC to fulfil their data release role. It is not appropriate to assume that departments currently have sufficient skillsets to successfully release data. Providing a checklist of actions will result in failure; instead the requirements point to a need for access to expertise in the field. For reasons of efficiency it could be argued that deploying such skillsets to each individual department is infeasible. Instead the NDC should be established as the data release hub. Departments would partner with the NDC to conduct a release, with the departments providing the domain specific expertise to understand the data, whilst the NDC provides privacy protection expertise to evaluate effective release strategies.

Release Process

The guidelines for publishing data should describe a process not a methodology. There are techniques that can be used for providing access to potentially sensitive data, for example, differential privacy and multiparty computation. The guidelines should describe a process which at the very least includes the following steps:

1. Understand the data – what does it show, what is the entropy, how could it be linked.
2. Create a risk register covering the potential release of the data – this will also determine who the data can be released to: public, semi-trusted, trusted, highly trusted.
3. Plan for deployment of mitigation strategies, with appropriate measurable evaluation (de-identification, differential privacy, multiparty computation).
4. Plan for breach management: should re-identification occur, consider the immediate steps to be taken to mitigate damage to individuals and organisations.
5. Decide and justify whether to proceed.

It is essential that the plan is constructed without prejudice and that concluding a dataset cannot be made available, or can only be made availability to highly trusted users, is a valid outcome. Once complete the plan should be published on the NDC website with a suitable period of time for public response and comment. The NDC should have the power to approve or reject a plan to release data.

We don't consider the decision on whether a dataset is available for public release or for limited sharing to be complex. If the dataset contains unit records about individuals it should not be publicly released. The challenge lies in determining what categories of trusted users should be able to apply for permission to access the data. This will require analysis of the sensitivity of the data and an understanding of the consequences of any potentially future data leak.

As discussed in the attached document focusing on data sharing, the public interest nature of datasets held by government is not necessarily present in commercial settings. As such the approach taken to commercial datasets, privacy and sharing should be distinct.

Conclusion - Going Forward

Different types of data deserve different treatments and different levels of protection. For example, data sharing for medical research may be acceptable (regardless of explicit consent) but data sharing with a person's insurance company is not. We welcome the draft report's trust framework, which appears to be an appropriate basic design for sending sensitive individual records to the researchers who need them, while publishing more generic government data openly. The decision about what should be shared under what conditions is a complex legal and political one.

Once it has been decided what should be shared and under what circumstances, implementing these decisions is a complex engineering problem. Indeed, it is often a collection of related but different engineering problems. Open publication of "de-identified" sensitive unit-level records does not seem to be the solution to any of these problems. Instead, a combination of secure storage, cryptography, random perturbation for differential privacy, and various other mathematical methods should be considered and applied to different datasets in keeping with different parts of the trust framework. This represents one of the hardest engineering design challenges of our time – it would be a serious mistake to expect a sequence of generic guidelines to solve the problem, when what's required is a major design effort from scientists and engineers who understand data analysis, cryptography and electronic privacy.

Confusing terminology prevents clear thinking. It is important to distinguish data that someone has tried to de-identify from data that actually satisfies the mathematical property of people being "not identifiable." Any standards or certifications for "de-identification" must distinguish very clearly between the two. Treatment involving removing obvious identifiers might be appropriate for secure storage with accredited scientists. This should not be confused with data in which people are "not identifiable," which is a necessity for sensitive data to be published openly.

Similar confusion befuddles the discussion of "open data", "public data", "government data" and "non-sensitive data." If one takes sensitive unit-level records and "de-identifies" them, this does not make the data "non-sensitive data" unless individuals are truly not identifiable. Data about government should indeed be open by default. Medical records, tax records, and other sensitive data about individuals should not.

The open government initiative should be an exemplar of its own principles. Government should be entirely transparent about its use of people's data, including down to the level of technical and mathematical detail. Citizens deserve to know what data government collects about them, how it is shared, when it may be published or linked, and how it will be kept secure. Openness is good for cybersecurity because problems and weaknesses can be found and fixed. This is an opportunity for Australia to set a high standard for protecting privacy, improving data access and protecting consumers.

Bibliography

[1] P. Ohm, “Broken promises of privacy: Responding to the surprising failure of anonymization.,” *UCLA law review 57,* p. 1701, 2012.

[2] Australian National Data Service, “De-identifying your data,” 8 December 2016. [Online]. Available: http://www.ands.org.au/working-with-data/sensitive-data/de-identifying-data. [Accessed 16 December 2016].

[3] “National Statement on Ethical Conduct in Human Research 2007 (Updated May 2015),” The National Health and Medical Research Council, the Australian Research Council and the Australian Vice-Chancellors’ Committee. Commonwealth of Australia, Canberra, 2015.

[4] D. Pauli, “Global 'terror database' World-Check leaked,” The Register, 29 June 2016. [Online]. Available: http://www.theregister.co.uk/2016/06/29/global_terror_database_worldcheck_leaked_online/. [Accessed 12 December 2016].

[5] P. Oborne, “HSBC, Muslims and Me,” BBC, 2 August 2015. [Online]. Available: http://www.bbc.co.uk/programmes/b0639w47. [Accessed 12 December 2016].

[6] N. S. a. B. Bryant, “VICE News Reveals the Terrorism Blacklist Secretly Wielding Power Over the Lives of Millions,” VICE News, 5 Feb 2016. [Online]. Available: https://news.vice.com/article/vice-news-reveals-the-terrorism-blacklist-secretly-wielding-power-over-the-lives-of-millions. [Accessed 12 December 2016].

[7] C. Duhigg, “How Companies Learn Your Secrets,” New York Times, 16 February 2012. [Online]. Available: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp. [Accessed 12 December 2016].

[8] A. Tanner, “How Data Brokers Make Money Off Your Medical Records,” Scientific American, 1 February 2016. [Online]. Available: https://www.scientificamerican.com/article/how-data-brokers-make-money-off-your-medical-records/. [Accessed 12 December 2016].

[9] A. Mahdawi, “Cookie monsters: why your browsing history could mean rip-off prices,” Guardian, 6 December 2016. [Online]. Available: https://www.theguardian.com/commentisfree/2016/dec/06/cookie-monsters-why-your-browsing-history-could-mean-rip-off-prices. [Accessed 12 December 2016].

[10] G. Petro, “Dynamic Pricing: Which Customers Are Worth The Most? Amazon, Delta Airlines And Staples Weigh In,” Forbes, 17 April 2015. [Online]. Available: http://www.forbes.com/sites/gregpetro/2015/04/17/dynamic-pricing-which-customers-are-worth-the-most-amazon-delta-airlines-and-staples-weigh-in/. [Accessed 12 December 2016].

[11] J. Sanburn, “Legal?, Delta Appeared to Overcharge Frequent Flyers for Weeks – Was That,” Time, 21 May 2012. [Online]. Available: http://business.time.com/2012/05/21/delta-overcharged-frequent-flyers-for-weeks-was-that-legal/. [Accessed 12 December 2016].

[12] J. S.-V. a. A. S. Jennifer Valentine-Devries, “Websites Vary Prices, Deals Based on Users' Information,” The Wall Street Journal, 24 December 2012. [Online]. Available: http://www.wsj.com/articles/SB10001424127887323777204578189391813881534. [Accessed 12 December 2016].

[13] D. Mattioli, “On Orbitz, Mac Users Steered to Pricier Hotels,” The Wall Street Journal, 23 August 2012. [Online]. Available: http://www.wsj.com/articles/SB10001424052702304458604577488822667325882. [Accessed 12 December 2016].

[14] ALRC, “Serious Invasions of Privacy in the Digital Era (DP 80) (Sect 12),” Commonwealth of Australia, 2014.

[15] “World-Check | Know Your Customer,” 12 December 2016. [Online]. Available: https://risk.thomsonreuters.com/en/products/world-check-know-your-customer.html. [Accessed 12 December 2016].

[16] P. Oborne, “Why did HSBC shut down bank accounts?,” BBC, 28 July 2015. [Online]. Available: http://www.bbc.com/news/magazine-33677946. [Accessed 12 December 2016].

Appendix

This appendix presents our responses regarding data sharing covered in the report, and in general. We have split this section from our main submission, because it contains a very brief overview and very few firm answers. Our purpose is to flag some of the complexities of protecting consumer privacy in an age when private companies can assemble vast databases about each individual consumer.

Data sharing is not restricted to just government data and the implications of lax rules on data sharing in the commercial sector present real risk of harm to the public. It is important that a strong framework is enacted to ensure that an inequality in access to information is not created between the public and the commercial organisations they interact with.

Although commercial entities have incentives to protect consumer privacy, driven by public relations and financial risk, they are ultimately driven by shareholder value and profit. Principles of government data sharing do not therefore apply immediately to the private sector.

Data Sharing in General

As mentioned in the summary of our main submission, the potential benefits from data sharing are significant; we support measures to streamline and improve access to datasets. However, policy makers must not lose sight of the competing motivations for data analysis. On the one hand we have public interest data analysis, which aims to facilitate data-driven policy. Access to such data should be streamlined and efficient; and the proposals in the report will contribute to achieving these goals. On the other hand commercial-driven interests aim to maximise profit. It is imperative that government and industry are not swept with the same public interest broad-brush.

The commercial sector should have access to non-confidential and non-personal data, as open data, as shown in Figure 9.1 of the report. However, as data becomes more sensitive, and potentially

identifiable, commercial access requires a high bar of justification. There will be cases where such access is appropriate for the public good, for example, in evaluating the effectiveness of pharmaceuticals. However, such access should be strictly controlled and limited to specific analysis and planned outcomes.

Chapter 4 of the report focusses on the potential for consumers to access and share the vast amounts of data collected about them by private corporations. This recommendation is predicated on the assumption that the collection and profiling of customers is beneficial. While such collection is (by definition) of benefit to corporations, it is not always good for the consumer; we discuss how the private collection of personal data can erode privacy and ultimately go against the public good.

Database of Ruin

Paul Ohm introduced the concept of the database of ruin in [1], in which he suggests that:

> *"…databases will grow to connect every individual to at least one closely guarded secret. This might be a secret about a medical condition, family history, or personal preference. It is a secret that, if revealed, would cause more than embarrassment or shame; it would lead to serious, concrete, devastating harm. And these companies are combining their data stores, which will give rise to a single, massive database. I call this the Database of Ruin. Once we have created this database, it is unlikely we will ever be able to tear it apart."*

It is essential that we do not inadvertently contribute to a database of ruin, and that we are proactive in preventing such a database's construction.

While such an eventuality might sound like conspiracy theory, there are already signs of a private database of ruin under construction, having a revealing tagline "Go beyond Know Your Customer" [5]. The Thomson Reuters World-Check database is widely believed to be one the largest database of politically exposed persons, and heightened risk individuals and organisations. A 2014 copy of the database was leaked in June 2016, revealing over 2.2 million records [4]. Previous reports have questioned the accuracy of the data [5] [6]. Sources of information World-Check have reportedly been from Wikipedia, blogs, and media organisations with ties to nation states [6]. In [6] World-Check is quoted as saying the following:

> *"We also provide secondary identifying information on individuals, such as dates and place of birth, and this will be similarly verified with reputable and official sources. If blog content appears, it is only as a supporting source for that secondary information, and is clearly identified as such."*

This clearly indicates a personally-identifiable database compiled with little or no permission or oversight from the individuals listed. Such databases are likely to proliferate without appropriate checks and balances.

Data Inequality

Risk of consumer exploitation is exacerbated by an inequality of access to data. Chapter 4 of the report includes the section "Consumer beware" that highlights the apparent paradox between consumers being concerned about privacy, whilst simultaneously engaging in actions that weaken privacy. Such paradoxical behaviour may suggest access asymmetry. Consumers are often unaware of the scale of the data that is being collected or the impact that data can have on their future interactions with businesses.

Corporations are readily building detailed, and at times, intrusive profiles of individuals [7] [8]. In [7] the New York Times describes the analysis Target undertook in the United States to profile their customers to find mothers in their second trimester – resulting in a high school girl receiving expectant mother marketing prior to her father even finding out she was pregnant.

The activities of companies like Target are rationally driven by profit. That is why, where inequalities or invasive techniques exist, market regulation is needed to protect consumers from exploitation. Today customer analytics is primarily used for targeted marketing. However, increasingly profiles are being used for dynamic pricing (sometimes referred to as Price Discrimination) [9] [10] [11] [12].

Dynamic pricing is a concept in which different customers will be charged differential rates depending on their profile and their likelihood/desire/necessity to complete a purchase. The pricing decision could be based on a profile [12] or on single session data [13]. In the case of [13], it was found that travel aggregator Orbitz was showing Apple Mac users higher priced hotel options, as a result of analysis that Apple Mac owners would spend more per night on hotels. The travel market is a good example of dynamic pricing and its potential usage. Online travel aggregators can provide consumers with a clear benefit: they can better choose service at an optimal price point. However, this situation rapidly changes when the companies we are purchasing from have knowledge of our actions. For example, imagine a scenario where a consumer is searching for flights and a hotel. They could be using a single price comparison site, or multiple independent sites. As the consumer conducts various searchers for hotels and flights the websites themselves, and potentially their partner advertising networks, will build a profile of what the consumer is doing and what they are likely to do. Where this becomes a problem is when the website or advertising network are able to predict the necessity of a prospective purchase, and therefore the amount the consumer is willing to pay. For example, when neither a hotel nor flight has been booked the prices are likely to be kept to a minimum to encourage custom. However, once the customer has paid for either the flight or hotel, their necessity for the other greatly increases. If the website knows this, it can increase its prices, confident that the consumer will pay a higher price because the cost of cancelling the flight outweighs the additional cost.

Our concern is not theoretical: there is both anecdotal evidence of dynamic pricing, as well as confirmed incidents across various countries and markets [12] [10]. The root cause of the problem is the inequality of access to data between the consumer and the provider. The consumer has no knowledge of how desperate the provider is to make a sale, they have no information about the capacity or current demand, and as such, have no way of leveraging that information to gain a better price. The provider, by contrast, has access to a detailed profile. "Taking your business elsewhere" – a common reaction – may not be effective. A number of websites will typically share an advertising partner, who may not explicitly provide information about past purchases, but could quite feasibly provide a metric indicating the likely demand for a service, based on a combined profile of the user.

Suggested counter measures often involve the equivalent of a digital disguise. Measures such as clearing cookies, browsing in incognito mode, usage of VPNs, etc, are suggested as ways to avoid tracking. It is easy to see that such measures would not be tolerated in the physical world. One would not tolerate being charged more than someone else for the same sandwich, just because they were hungrier. Likewise, if someone suggested they don a disguise and return later in order to purchase the sandwich at the same price it would be considered ridiculous. Yet we tolerate such behaviour in the online data driven world, not through choice, but through necessity due to inequality.

Businesses will tend not to be transparent about customer profiling or dynamic pricing. Confirmation of such strategies could significantly reduce consumer goodwill. The report asserts that "Commercial entities will in most cases have a clear financial incentive to ensure that the privacy of their customers is protected," but in practice incentives to protect customer privacy are not always sufficient or even present. One motivation for keeping data private, from both customers and competitors, is competitive advantage.

Data inequality does not only exist in the marketplace, it is also evident in the law. The act of incorporation creates a legal entity with some of the same rights as a physical person. Such an entity is sometimes referred to as a juristic person or, in the United States, as corporate personhood. Corporations do not have all the rights of a physical person, for example, they cannot marry, and as such, one would expect them to have a subset of the rights of a physical person. However, corporations have traditionally had a greater degree of protection of their private information, via protections of trade secrets and breach of confidence. That situation may change in the future, with suggestions that breach of confidence could be used to protect individual privacy [14]. However, as it stands the inequality is striking, and often enshrined in the terms of services that create a binding agreement between a consumer and a website. For example, on expdia.com.au the terms of service include the following statement:

> *"…you agree not to modify, copy, distribute, transmit, display, perform, reproduce, publish, licence, create derivative works from, transfer, or sell or re-sell any information, software, products, or services obtained from this Website"*

which is followed by a requirement for the consumer to agree to not:

> *"access, monitor or copy any content or information on this Website using any robot, spider, scraper, program, algorithm or other automated means or any manual process for any purpose without our express written permission;"*

These clauses effectively forbid the very actions that the website will undertake with the personal information acquired from the user's interaction or transaction. The first would prevent the collection, construction and sale of detailed customer profiles. The second would prevent analysis of a customer's data and actions. The second clause could even be considered as forbidding a consumer manually monitoring prices. These terms of service are rarely, if ever, tested in court. However, they provide a useful insight into the perspective of the corporation, and are effective at highlighting the inequality in terms of access and use of data.

One potential solution is for information supplied by individuals to corporations, for the purposes of conducting business, to be considered commercial in confidence by default. This could prevent consumers becoming a secondary product, whose data can be analysed and sold. Privacy policies and terms and conditions should not override the implicit commercial in confidence status of their data. If a company wished to collect, analyse and potentially sell such information they would need to explicitly collect it for that purpose, and not derive it from information collected as part of an unrelated transaction.

Right to Access

Recommendation 9.2 discusses the rights to access and correct data held by organisations. Such a right is essential in both holding to account entities collecting data and ensuring the accuracy and integrity of the data. We fully support such a recommendation, but would caution that the process for making data digitally available to a consumer must be appropriately secured. It is an important

part of the rights a consumer should have, but it is also imperative that it does not become a new attack vector which miscreants can use to acquire personal information.

We caution against imposing charges for correcting errors in data. It should be the responsibility of the data holder to ensure accurate data, and the cost burden for doing so should be met by the data holder. If they do not wish to incur that burden they should not hold the data. To place a cost on correcting errors disproportionately penalises lower income groups and could create a situation where a disreputable organisation deliberately corrupts data in order to derive income from corrections.

The recommendation currently does not permit a consumer to remove their data from a data holder. Clearly such a provision is going to be challenging to implement, and would require careful consideration. However, there is some merit for considering such a policy. It would be easier to implement were our suggestion for consumer data to be treated as commercial in confidence, by default, be adopted. In such circumstances when a consumer terminates their interaction with a supplier they would have the right and expectation that the data shared during the interaction would cease to be used. It would be feasible for the data holder to delete such data from its active database. There would need to be exclusion for requiring removal from backups, a common clause in commercial contracts already. A key reason for deletion of such data is that without it a consumer cannot be proactive in protecting their data. For example, if a data holder suffers an unrelated breach, or is found to be deficient in its data protections, a consumer can do nothing to mitigate the future risk or hold the data holder to account. Additionally, where businesses are taken over, it is perfectly legitimate for a consumer to wish to sever ties and disallow unintended data sharing.

Conclusion

As decisions are made about the future data infrastructure in Australia it is imperative that the privacy of individuals be protected. We support the recommended right to access, and further argue that a right to remove (delete) data is essential to hold business to account and to allow consumers to be proactive in protecting their privacy. Once lost, privacy is extremely difficult to regain. Globally we are seeing an ever expanding industry built on profiling and analysis of individual consumers with detrimental effects. Regulation around data sharing for public good vs profit should be differentiated. Profit-making endeavours that drive economic prosperity should be supported, while making the restrictions, rights and protections afforded to consumers and their data stronger and more explicit. If we fail to redress the data inequality that currently exists, we run the risk of creating a database of ruin.

Summary of recommendations:

- For designing data sharing policy, private entities should not be assumed to be operating in the public interest. Their access to datasets, and indirectly the data we make public, should be managed to avoid disadvantage to the general public.
- Introduce (or retain) restrictions on potentially damaging personally-identifiable databases constructed by private corporations, which could otherwise lead to a ‘database of ruin’.
- Consumer data should be treated as commercial in confidence when provided within a transaction. When a consumer buys a product or service their data should not become a secondary product for a business to harvest and sell.
- Consumers should be better educated on the risks to their privacy by data collection and encouraged to be wary of engaging in actions that undermine their privacy.