Submission to Senate Standing Committees on Legal and Constitutional Affairs in relation to the Privacy Amendment (Re-identification Offence) Bill 2016

10 December 2016

1 Introduction

This submission addresses perceived issues in the draft of the Privacy Amendment (Re-identification Offence) Bill 2016, referred to in this submission as the "proposed law".

I work as a Data Analyst, and have a PhD that focused on cross-matching scientific datasets. I also have an interest in computer security from a personal perspective, focusing on how I can securely interact with the internet and protect the personal information of me and my family from disclosure. From my background working regularly with large datasets, including those published by the ABS and other government agencies, the ease and speed with which any size dataset can be transformed and distributed, without requiring any resources past those provided by a commercially available cloud service, is evident to me.

It is my view that the proposed law, as tabled in the Senate for discussion, needs significant amendments. In its current form it will not effectively prevent either datasets from being reidentified or reidentified datasets being distributed. It introduces unnecessary legal threats for law-abiding citizens wanting in good faith to correctly and promptly report issues with "deidentified datasets", including removing the presumption of innocence by reversing the onus of proof and not including exemptions for researchers or good faith discreet reporting in the proposed law.

2 Agencies not responsible for publishing flawed datasets

The proposed law relies on the assumption that datasets containing aggregated or obfuscated personal information are able to be legitimately and unquestionably published solely on the "basis that they are deidentified". This gives the impression that the datasets and the agencies which publish them, do not need to be examined as part of determining whether a reidentification offense has taken place. Some datasets which are published as deidentified, leave non-trivial identifiable information intact, whether it is in the strict class of "personal information" or not, and are hence trivially reidentifiable to some degree. These datasets should never have been published in the first place. However, the proposed law does not refer to publishers at all, and refers in the Explanatory Memorandum to other possible places where it may be dealt with for the "personal information" case, but not for other cases where non-personal information can be used to enable reidentification.

The assumption that datasets may be safely published and labelled as "deidentified" is flawed on the basis that it may be an undefinable concept in an environment where all information is digitally available and the integration and reuse of datasets is encouraged to derive new value. The proposed law does not refer to a particular likelihood of identification by a normal user of the dataset as part of routine dataset integration operations and what level of resources or other datasets may be required for reidentification before it is a legal routine operation and not an illegal "reidentification act".

For example, would a reidentification offence be committed if someone was identified to live in a particular suburb, or street, or a particular ABS Mesh Block (approximately 30-60 dwellings) [1], without knowing their name or other personal information protected by the Privacy Act? Is it also a reidentification offense to either directly publish or reidentify a single household or a small group down to the suburb/street/Mesh Block level.

The Office of the Australian Information Commissioner (OAIC) has published a guide for deidentification that refers extensively to risk assessment and risk management, which are vague concepts if not backed up by specific guidelines [2]. However, it does not recommend or prescribe any standard method to follow, referring instead to a hypothetical "motivated intruder" test, using a definition that does not fit either the "motivated" or "intruder" terms as they are used in security research. Both this and the other recommended test ("in the round") assume that personal information available on the internet will be few and far between, and there is a non-trivial level of difficulty to integrate it. Data integration is a fairly widely available skill, with many freely available tools published by large software companies such as Google and Facebook. In current academic society as encouraged by government policies that encourage data reuse for innovation, both tests are and will be redundant, even with the chilling effect introduced by the proposed law.

In another example, the Statistical Linkage Key (SLK) method [3] that the ABS may or may not be using internally or in communications with other agencies to obfuscate census data, is trivially reidentifiable. Every SLK contains, and makes no legitimate attempt to cryptographically hide, the personal information it may be aiming to deidentify through its basic obfuscation technique. Hence, any agency that publishes or transmits SLK identified records, even if they claim they are doing so "on the basis of deidentification", is publishing records that are trivially reidentifiable and should be stopped before the datasets become accessible through a data breach or an authorised publication.

The proposed law should be amended to make it an offence for agencies to leave any information that could potentially be used for reidentification in datasets that are published including those published "on the basis that they are deidentified", or otherwise. The proposed law aims to make a dataset user justify all of their actions starting from the point that they are guilty, shifting the blame for the dataset being incorrectly published from the dataset producer to the user. By not making it a specific offense for an agency to insufficiently deidentify datasets, the proposed law is targeting users for mistakes that agencies have made.

3 Reidentifiying own personal information

The recent Medicare dataset reidentification issues [4, 5] in which doctors were able to be reidentified (which may have been a factor in this law being proposed given the timing of the initial press conference), may have made it necessary to prosecute doctors for identifying themselves in deidentified datasets. This is despite the cause being that Medicare did not adequately deidentify the data that doctors submitted to Medicare. Doctors submitted the personal information to Medicare in good faith, and on the understanding that it would be held securely and used appropriately without violating their patient confidentiality requirements. Even if the doctor had notified Medicare as soon as they became aware of the issue and not publicly disclosed or discussed it with other affected individuals, under the proposed law they could still have been guilty of the criminal provisions for simply identifying themselves, while the agency would not be held accountable for any of their actions.

The proposed law must be amended to give immunity to anyone for reidentifying their own personal information, or one of their dependents or clients, in a published dataset, whether it has been labelled as deidentified or not, regardless of the way notifications to agencies occurred. If someone recognises their own personal information, or that of a child or other dependent, has been publicised, they should be free in a civil law-abiding society to report it, using whatever channels they are aware of, without consequences. The current text of the law explicitly does not assure them of the freedom to notify an agency without facing criminal provisions if they use the wrong channels or they fail to follow the other provisions which they are unlikely to be aware of, and would need to defend themselves against starting at the position of being guilty.

4 Reporting issues to agencies

It is not always obvious who to report issues with published datasets provided by agencies to, and it is very rare that reporting can be done securely to minimise the risk of anyone who is not authorised to see the report observing it. This lack of a quick, confidential channel will prevent legitimate reidentification notifications from occurring unless the proposed law is amended to require a publicised notification channel. In addition, there is no clearly visible method of requesting that a published dataset be returned and copies of it be destroyed in cases where reidentification issues have been identified.

The proposed law should be amended to require agencies to establish and maintain an authoritative, secure, and anonymous channel for notifications to encourage dataset users to promptly notify agencies about issues with their published "deidentified" datasets, including those that may not have been labelled as deidentified but which erroneously contain private personal information. An example of a system which may provide an anonymous channel for notifications is the SecureDrop system created and used by journalists to communicate securely. [6] If the core purpose of the law is to maintain the confidentiality of personal information, as expressed in the Explanatory Memorandum, then there should be no complaints about secure anonymous reporting that enables prompt notification to occur to minimise the time that flawed datasets are publicly available.

5 No guaranteed exemptions from Parliament, only from Executive

The proposed law does not specifically allow for exemption in any cases without ministerial decree for discreet (not public) notification to agencies that they may have published datasets that rely on flawed deidentification procedures. All notifications carry with them a criminal and civil penalty unless the notifier knows that they are specifically exempt. Given the time sensitive nature of notifications, it may take too long for a citizen, of whom none are automatically exempt, to identify their exemption status compared to the time available to prevent the flawed dataset propagating further across the internet.

For example, in the recent Australian Red Cross blood donation data breach [7], the security researcher who found that a pristine copy of a donation records database had been accidentally published on a website was not confident to disclose the issue directly to the Australian Red Cross, possibly due to the lack of protection given to security researchers by Australian Law. They instead privately notified a popular Australian security researcher, Troy Hunt, who runs a website, `https://`

haveibeenpwned.com/, which makes it easy for people to discover if their personal data has been leaked in data breaches. Hunt was personally confident, and has a public track record, of being able to discreetly follow up data breaches in an ethical manner. Similar issues will also occur for reidentified datasets if exemptions are not included in the law to allow to give immunity for good faith discreet notifications of both general data breaches and reidentification issues with datasets.

There is no evidence that there is an underground industry dealing specifically in reidentified datasets. There is however evidence that there is an underground industry dealing in breached datasets containing unredacted personal information [7]. That industry will not be hampered by further chilling the community of ethical security researchers looking for issues in intentionally published datasets to protect themselves and other citizens from Government mistakes. It is self-evident that in both cases, data breaches and reidentification, the faster notifications to agencies proceed, the faster datasets can be removed from public view and fixed. The proposed law will slow the process of notification by adding unnecessary penalties for those who become aware of issues if they are not decreed to be in the currently declared list of exempt individuals or classes, if Parliament does not include a base list of classes by default.

The proposed law should be amended to directly include guaranteed classes of citizens who are exempt, from both reidentification and notification offenses, as long as they are acting in good faith, so that the Courts have a basis from Parliament, on which to make exemption decisions without relying on being directed by the Executive at some time in the future. A shortlist of classes that the Minister desires to use as a basis for exemptions has already been published in the Explanatory Memorandum, and there is no reason not to include all of those in the proposed law itself. In conjunction with the backdating provision, the lack of exemptions has already caused a security researchers to question the usefulness of the law in effectively targeting those who may be performing reidentification activities in bad faith.

6 Removing presumption of innocence

The proposed law reverses the onus of proof, making it necessary for a researcher to argue in Court that they are exempt, which requires knowledge of the current list of exemptions, and is made more difficult given that the proposed law contains no exemptions for Parliament to debate and vote on before approving the proposed law. This is unreasonable and there is no evidence provided that the government prosecutor would obtain more rightful convictions by requiring that a researcher defend their inclusion in an exemption decree.

Together, the onus provisions, and the lack of exemptions by default, will result in attempts by cryptology researchers to use their tools to obfuscate themselves as the source of legitimate discreet notifications, or they will forward their findings onto researchers who are aware of their exemption status and thus able to confidently disclose issues to agencies without being in fear of criminal prosecutions themselves [7].

There is no evidence that criminal provisions hamper unethical security researchers, as they, more than anyone, know how to anonymously republish and propagate datasets and without having themselves identified as the source. However, badly designed provisions including notification penalties and self-incrimination in relation to reidentification, such as the proposed law contains, will have a chilling effect on ethical researchers who are not looking to publish, and will be affected by the proposed law for discovering and reporting issues if they are not in one of the ministerially declared exemption classes at some point in the future.

The proposed law should be amended to not reverse the onus of proof. The Explanatory Memorandum makes it clear that it should be trivial, so the prosecutor will be able to prove it easily themselves and not cast dispersions on ethical security researchers in regard to a presumption of guilt before trial.

The proposed law should be amended to introduce the requirement that the prosecutor prove that prompt good faith discreet notifications did not take place before applying any of the criminal or civil penalties. Leaving the onus on researchers to defend themselves from a starting position of being guilty, encourages them to stay silent rather than promptly help agencies fix their mistakes to protect the community from badly published datasets.

7 Backdating the law

The suggestion that lawfully abiding citizens will be using the time while Parliament debates this law to reidentify datasets without prosecution is farcical, and sets an unnecessary precedent for backdating laws to a press release date, when the press release did not contain any significant details about the law in question and occurred well before the date on which the proposed law was presented to Parliament.

The proposed law should only be valid from the day it is given assent by the Governor General, to provide surety for the large community of ethical security researchers that they will not be prosecuted using a law that was not available on the date they allegedly violated it. In combination with all exemptions being declarations, which could be reversed, backdating has not provided any surety to researchers that they are not the main target of the law.

8 Parliament should be able to reverse controversial decisions made by the Executive

If the proposed law was amended to include a base list of classes of exempt citizens, then further declarations would not need to be exempt from section 42 (disallowance) of the Legislation Act 2003, as the majority of classes of citizens affected by the exemptions would already have certainty and not need to be provided temporary assurance by the potentially partisan Executive who could still change the list at will. The general public do not trust the Executive more than either the Parliament or the Courts. By not enabling the Parliament to overrule unfair decisions by the Executive in relation to exemptions, the law reduces the checks and balances that the Parliament and the Courts are able to provide in our civil society.

The proposed law should be amended to enable the Senate to disallow controversial decisions made by the Minister in regards to exemptions.

9 Consultation

As the current Attorney General, who has proposed this law to Parliament, has recently eloquated in his evidence given to this committee, the term "consult" must be taken with a grain of salt. In particular, the actual subject matter does not need to be discussed at a meeting or in follow-up discussions to fit the Minister's preferred definition of "consult". Hence, the provision in 16G(4) of the proposed law, that the Minister "must consult the Commissioner" is rendered worthless in terms of its apparent goal of giving the appearance that it is not merely the Executive making arbitrary decree's in relation to exemptions to the proposed law in future. Along with the lack of notes taken at previous "consultations" by the Attorney General in relation to other legislative instruments, it has not encouraged the security research community that the Minister may have the requirement to consult with others in relation to exemptions for reidentification offenses.

The proposed law should be amended to include a shortlist of necessary discussion points for consultations with the Commissioner, a permanent record of the agenda and decisions taken, and a requirement that the Commissioner agree with each specific exemption. This would give a permanent record that would reflect the decisions taken for each particular case, to give surety that process is being followed for Ministerial exemptions to criminal offences.

10 Human Rights

The Human Rights justification given in the Explanatory Memorandum effectively argues that individual rights of good faith researchers are dispensable and that the use of criminal prosecution to create fear for legitimate citizens without any inbuilt protections or defenses fits with international human rights treaties. In doing so, it confuses the separate issues of prevention and prosecution. A preventative solution would emphasise that agencies are solely liable for releasing poorly deidentified datasets, as the alternative, strongly deidentified datasets, cannot reasonably be expected to ever be subjected to the provisions in the proposed law. The strategy espoused in the proposed law, to solely prosecute dataset users, criminalises good faith security researchers and ignores the causes of why poorly deidentified datasets are being released, and will continue to be released in the future.

There are other more pressing issues in regards to protection of personal information to preserve basic human rights in relation to digital services. As highlighted by the Australian Red Cross blood donation data breach [4] and others, there are much larger instances of human rights violations from cases where entire pristine databases have been released to the public without any obfuscation or deidentification. The status of privacy law in relation to mandatory data breach notification to both the government and to users in these cases should be a higher priority than the ill-defined issues in the proposed law concerning reidentification of intentionally published datasets.

Another example of a case where human rights in relation to pristine private information are not well protected currently is in the continued use of the insecure-by-design Signalling System No. 7 (SS7) international phone system [8], along with the Short Message Service (SMS), as the basis for Two Factor Authentication (2FA) systems. SMS is used as a 2FA method for the Governments centralised `my.gov.au` portal, which it a much larger target for the inappropriate release of private personal information, compared to the 10% Medicare deidentified dataset, particularly as a citizen can access their entire identified Medicare and Australian Tax Office histories through `https://my.gov.au`. SS7 is known to be able to be hacked at will by any country participating in the system, to intrude on the human rights of any citizen in other countries that utilise SS7-based phone networks, making my.gov.au a potential target that is much more pressing, given the lack of preventative measures in the proposed law. The `my.gov.au` portal should move to a more secure 2FA system such as One-Time Password (OTP) mobile applications or tokens for authentication, to prevent the leakage of Medicare/ATO/Centrelink data.

The proposed law should be amended to explicitly note that users are not responsible for failures by agencies to correctly deidentify datasets that they publish or protect deidentified datasets that are leaked before publication.

References

[1] http://www.abs.gov.au/ausstats/abs@.nsf/0/A53A152BBF2992EBCA257801000C64BE

[2] https://www.oaic.gov.au/agencies-and-organisations/business-resources/privacy-business-resource-4-de-identification-of-data-and-information

[3] http://www.abs.gov.au/AUSSTATS/abs@.nsf/lookup/4240.0.55.002Chapter15022011

[4] http://www.abc.net.au/news/2016-09-29/medicare-pbs-dataset-pulled-over-encryption-concerns/7888686

[5] https://data.gov.au/dataset/mbs-sample-10pct-1984-gz

[6] https://securedrop.org/

[7] https://www.troyhunt.com/the-red-cross-blood-service-australias-largest-ever-leak-of-personal-data/

[8] https://en.wikipedia.org/wiki/Signalling_System_No._7#Protocol_security_vulnerabilities