Comments on the proposed amendments to the Privacy Act to criminalise re-identification

Chris Culnane, Benjamin Rubinstein, Vanessa Teague
University of Melbourne
November 2016

In September we discovered that the method of encryption used in the MBS/PBS linked 10% sample dataset was insecure. We showed that it was possible to reverse the encryption to re-identify the MBS suppliers listed in the dataset. We notified the Department of Health, who took the data down from the data.gov.au website. It is good for privacy and cybersecurity when errors and weaknesses are discovered and fixed.

The plan to criminalize re-identification was announced just before the public announcement of the problem in the MBS/PBS encryption. The amendment seems intended "to improve protections of anonymised datasets." [1] This may be a valid goal for *legal* protections, but the effect on the technical protections could be exactly the opposite. The threat of criminal penalties could inhibit open investigation, which could mean that fewer Australian security researchers find problems and notify the government. Criminals and foreign spy agencies will be more likely to find them first.

We encourage the committee to consider what the Bill's objectives are and whether it meets them.

Does it prohibit all uses of de-identified government data to do harm?

Does it prohibit some cases of re-identification that are harmless, are legitimate contributions to public debate or scientific research, or would have a positive effect on privacy by identifying and correcting problems?

A key theme of our questions is to distinguish between re-identification *per se*, and the use of de-identified government data to do harm. The two are not the same. We believe that re-identification should not be a crime, though some uses of government data should be.

We begin with some imaginary scenarios to draw the right distinctions, then offer some specific questions and suggestions about the proposed amendments. In each of the examples below, consider whether the behaviour does harm or not, and whether it would be prevented by the proposed amendments or not.

[1] Senator the Hon George Brandis, "Amendment to the Privacy Act to further protect de-identified data", Media Release, 28 September 2016.

Scenario A: A privacy researcher uses legally available and public information to recover a state Premier's medical record from a de-identified dataset. She prints it out and mails it to him with a polite cover letter suggesting that the state should improve privacy protections.[2]

Scenario B: An Australian journalist uses published government data to re-identify someone (Person B) and hence learn that he is HIV positive.
B1: The journalist writes an article exposing Person B's HIV status in an effort to embarrass him.
B2: The journalist obtains permission from Person B and uses his case as an example in an article criticising government protection of personal data.

Scenario C: The same as B, but without re-identification. For example, the journalist might infer Person B's HIV status by narrowing him down to two de-identified records, both being HIV positive.

Scenario D: A bank uses credit card billing data to link its customer records to a published government dataset.
D1: The bank identifies particularly healthy customers and offers them discounted life insurance.
D2: The bank identifies customers with a terminal illness and cuts off their line of credit.

Scenario E: The same as D, but without explicit re-identification of the government data. For example, the bank could remove the names from its own database before linking, mark the record with a decision about whether to offer life insurance or cut off credit, and then link the decision back into the bank's database without explicitly matching names to the government data.

Scenario F: A man with mental health problems is concerned that his employer might re-identify his record in a published government dataset by using the dates of his submitted medical certificates. He wants to know whether his record is included and is reasonably identifiable in this way, so he asks his wife to query the relevant dates in the dataset. She retrieves a unique record, which they can confirm is his.

We suggest rethinking the Bill to focus on what does harm (and is not already illegal) rather than on re-identification itself.

What does “de-identified” mean?

The proposed amendments apply to ‘de-identified personal information.’ The Privacy Act does not define ‘de-identified,’ but the definition of ‘personal information’ is met if there is ‘an individual who is reasonably identifiable’ from the information. Successful re-identification is a

2 The other examples are invented, but this example actually happened. The state was Massachusetts, the researcher (Latanya Sweeney) got a job at Harvard, and the government stopped releasing "de-identified" medical records.

demonstration that some individuals are reasonably identifiable. We are not sure what the definition of 'de-identified personal information' is – mathematically, if it can be re-identified then it was not properly de-identified.

The proposed criminal offences apply if "the information was published on the basis that it was de-identified personal information" 16(D) 1(b). This seems to refer to the intentions and beliefs of the publisher, rather than to mathematical facts about the difficulty of re-identification.

If a dataset was published in the belief that it had been securely de-identified, but actually it was easy to re-identify people, would the government want to find out? Who would tell them?

Why open investigation is good for security

The threat of jail time discourages law-abiding Australian researchers and journalists from making the simplest and most convincing demonstration that a de-identification method has failed. If the new rules had been in place in September, we would not have discovered the problem in the MBS/PBS dataset encryption, the dataset would probably still be up, and the government could be unaware it was insecure.

There is nothing wrong with employing experts to examine datasets *in addition* to free research. However, prohibiting open scrutiny simply cuts off the best way of learning about mistakes. In election software, we have found errors and security holes in software that had been certified by experts engaged by the electoral commissions. A thriving community of free cybersecurity researchers is bound to discover some problems that a single designated expert missed.

There is some ambiguity over exactly who is affected by the Bill. It seems that most Australian universities are not covered by the Privacy Act, though the implications are unclear for students, ANU researchers, and individuals acting on their own initiative who happen to be university employees. An explicit exemption for scientific research or public interest would help to reduce this confusion.

Legislation cannot make something secure or prevent criminals or foreign groups from exploiting its weaknesses – it can only stop Australians from pointing the problems out. The more people who can investigate legally, the higher the likelihood that the government learns about a problem before criminals do. Decisions about what de-identification methods to use, and which datasets to publish, could then be made based on grounded mathematical facts. Some companies (such as Google[3]) actively encourage this kind of investigation and pay rewards for responsible notification of successful exploits. This does not guarantee that everything will be perfectly secure or that all government decisions will be well-informed, but we are unaware of any better way to approach either of these objectives.

3 https://www.google.com/about/appsecurity/chrome-rewards/index.html

The pitfalls of retrospective legislation

The question of exempting legitimate research was raised immediately when the amendment was first proposed. A spokesperson for the Attorney General's office told the media the day after its first announcement that "There will be provision made for legitimate research to continue." We interpreted this to mean there was a commitment that *all* legitimate research would be allowed to continue. When the draft Bill was released, the only research exemptions were by explicit ministerial determination or by contracting with the relevant agency. At the time, we interpreted this as a deliberate backflip that put us at risk of a serious criminal penalty. We now realise that the Bill was entirely consistent with the original announcement, but not with the *all* that we had assumed to be implicit.

Whether you agree that *all* legitimate research should be exempt, or prefer that only *some* designated research should be exempt, we hope it is clear that ambiguity over what does and does not incur a prison sentence is highly undesirable, particularly for a retrospective law. Researchers might well be left in the ridiculous situation of being unable to tell the government what they had discovered during the time that they thought the investigation was legal, for fear of going to jail over a misunderstanding.

A note about responsible disclosure

We suggest that responsible disclosure rules are expanded to include notification to other appropriate authorities such as CERT Australia or the Australian Information or Privacy Commissioner. Not every government department even has a way of receiving this sort of notification, nor is it always obvious which department is responsible for any particular dataset. The Department of Health were entirely rational and responsible in their response to the news about the encryption error, but our prior experiences with other authorities have taught us that not everyone responds constructively to a notification of their own mistakes. It is important that someone who learns of a serious problem can report it to an appropriate authority who is not the responsible department.

Summary

Although we agree that some uses of re-identified or incompletely de-identified data should be prohibited, we see no good reason to prohibit re-identification itself.

Criminalizing re-identification without a clear and explicit exemption for research or a defence on the grounds of public interest will be bad for privacy and information security. It will make the government far less likely to learn about a problem before criminals and foreign governments do.

The best way to improve protections of anonymised datasets is to permit free and open re-identification combined with responsible disclosure.