Cloud Insurance and the University of Technology Sydney

Data Protection and opting out: a snapshot of international developments being driven by the United Nations Special Rapporteur on the Right to Privacy

In 2015, the UN appointed a Special Rapporteur on the Right to Privacy. The appointment was in response to the Snowden allegations and work that was undertaken in their aftermath by the Human Rights Council.

The SRP produced his first report in March 2016. The report identified a number of key themes that require investigatory work under the SRP's mandate. One of these is Big Data and Open Data.

In July 2016, at the SRP's Conference on Privacy, Personality and Information flows at the New York University law school, I was appointed to lead this part of the SRP's mandate. My key task is to oversee and coordinate production of a paper on the privacy implications of big data and open data for presentation to the UN General Assembly and the UN Human Rights Council in late 2017.

Big Data is a major topic of discussion across a range of disciplines, including law enforcement, national security, marketing and advertising and scientific research. Although it has dropped off the Gartner hype cycle, having passed the 'Peak of Inflated Expectations' it is now considered to have become so ubiquitous that it has been incorporated into many other hype cycles.

There are no accepted definitions of big data: there are only descriptions:

> Big Data refers to the inability of traditional data architectures to efficiently handle the new datasets. Characteristics of Big Data that force new architectures are:
>
> - ***Volume*** (i.e., the size of the dataset);
> - ***Variety*** (i.e., data from multiple repositories, domains, or types);
> - ***Velocity*** (i.e., rate of flow); and

- ***Variability*** (i.e., the change in other characteristics). [1]

The NIST description lists four Vs but there are others that are commonly used. The lack of a definition poses a number of conceptual problems for the big data theme – how do you go about determining risk when you don't really know what you are measuring or assessing? To illustrate the problem, it's worthwhile taking an historical perspective. For example, the Doomesday Book, compiled in 1086 as a survey of land and chattels over the whole of England, must fall within eleventh century experience as the Big Data of its day. So too the inventory of the English abbeys undertaken by Thomas Cromwell in 1536. So too the 1933 Prussian census that used the Hollerith punch cards and computing machines supplied and maintained by IBM that produced the evidence of religion that underpinned the Holocaust. The big data of today can easily become the little data of tomorrow.

Open Data can be seen as one of the dimensions of big data - as an input or data source. It is defined as 'data that can be freely used, re-used and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike.'[2]

Open data has become a public sector article of faith over the last few years. The asserted policy basis for this is that 'governments have a significant amount of data that can be published publicly. Where this data is made available in a machine-readable way, digital services can leverage it to support improved information and service delivery for users.'[3]

The SRP has expressed reservations about Open Data:

> At first sight Open Data sounds fine as a concept, a noble and altruistic approach to dealing with data as a common good, if not quite "common heritage of mankind". Who could object to data sets being used and re-used in order to benefit various parts of society and eventually hopefully all of humanity? It is what you can do with Open Data that is of concern, especially when you deploy the power of Big Data analytical methods on the data sets which may have been made publicly available thanks to Open Data policies.

[1] See NIST, http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

[2] See http://opendatahandbook.org/guide/en/what-is-open-data/

[3] Digital Transformation Office, https://www.dto.gov.au/standard/design-guides/open-data/

There are now a significant number of data sets that have been released by government in Australia under Open Data policies. One of the most recent and significant was the release of more than 1 billion lines of what is claimed to be de-identified historical health data by the Department of Health. The Department stated that:

> To ensure that personal details cannot be derived from this data, a suite of confidentiality measures including encryption, perturbation and exclusion of rare events has been applied. This will safeguard personal health information and ensure that patients and providers cannot be re-identified.[4]

There is no doubt that better research for the common good, for example population health research, carries with it community benefits. But what are open data's privacy risks and can they be mitigated appropriately? It's interesting that the Department's announcement did not include the specific details of the nature of the de-identification process it used.

The SRP's Big Data and Open Data theme

The big data/open data theme has been divided into a number of areas of inquiry. These include:

- The benefits of big data and open data
- The associated data protection risks
- The ways that the risks can be managed mitigated

The focus in this presentation is on one aspect of risk mitigation, privacy enhancing technologies or PET.

Usually PET is used to refer to the use of technology to help achieve compliance with

[4] http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/News-20160811-10PercentSample

data protection legislation. The rationale for using PETs does not end with privacy. PETs can protect corporate confidential information and intellectual property as well as other categories of valuable information.

De-identification

One of the main PETs is de-identification. Privacy law only applies to personal information. If the information no longer falls within the definition, privacy no longer applies.

De-identification is one of the most contentious and hotly debated international privacy issues. Its supporters acknowledge that even though no de-identification approach can be guaranteed to be successful all of the time and for all time, robust and risk based de-identification processes can provide sufficient protection to comply with privacy laws. They argue that there are no guarantees of anything. Ann Cavoukian and Brian Castro are two of the most prominent supporters of de-identification.[5] Opponents of de-identification argue that (i) there is no evidence that de-identification works either in theory or in practice and (ii) attempts to quantify its efficacy are unscientific and promote a false sense of security by assuming unrealistic, artificially constrained models of what an adversary might do.[6]

At this stage its difficult to know who is right and who isn't, or whether a binary answer to the de-identification debate is either helpful or useful. Perhaps we need to look at the debate in a more nuanced way, accepting that in some, but not all cases, de-identification might provide acceptable answers. But even so, it's difficult to see where the boundaries lie. Its becoming easier to combine de-identified data with other data sources in ways that increase the risk of re-identification.

As a skeptical lawyer, I have reservations about relying on any one technology as a

[5] See Ann Cavoukian & Daniel Castro, Big Data and Innovation, Setting the Record Straight: De-identification Does Work (2014), available at http://www2.itif.org/2014-big-data-deidentification.pdf
[6] See http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf

permanent solution to de-identification issues. Technology changes too quickly to provide a failsafe, permanent foundation for protecting rights.

Distributed ledgers

High on the hype cycle is distributed ledger technology of which block chain technology is a component. A distributed ledger 'is a consensus of replicated, shared, and synchronized digital data geographically spread across multiple sites, countries, and/or institutions.[7] It is a database spread across multiple sites, nations and institutions. Data is stored in a continuous ledger but can only be added when the participants reach a validation consensus. More about the validation consensus process is discussed below.

A block chain takes data or records and stores them in a block. A simple analogy is the recording of a transaction on a piece of paper. Each block is then 'chained' to the subsequent block using cryptography. The chain of blocks becomes the digital version of a ledger. The ledger can be shared and examined by anyone permitted to do so. The key difference between this process and a conventional database is that rules can be set at the transactional level in the block chain whereas this does not occur with conventional databases.

In a short presentation it is impossible to canvass all of the viewpoints about distributed ledgers and block chains. We are in the midst of an incredible explosion of information about the uses of these technologies: experience suggests that we need to carefully examine and understand their strengths and weaknesses before reaching firm conclusions about their effectiveness.

But we need to note a few issues about privacy impacts and risks at the outset. Block chains offer new opportunities for individuals to collaborate to create datasets in a peer network without a central intermediary. But what if the ledger is

[7] See http://www.blockchaintechnologies.com/blockchain-definition

controlled by a single entity or a group of affiliated interests who control the validation and permissions process? What happens if they exercise their majority powers? Further, what if the data in each block contains personal information – such as your health records, or your recent bankruptcy, or your change of gender? In an open block chain, chances are this information is available to anyone and forever. In a closed block chain it is open to those with the relevant permissions. Although encryption can be used to protect personal information embodied in each block, at some point the permissions and validation processes require it to be decrypted.

The distributed ledger technology model operates in a way that challenges one of the main information privacy assumptions – that *organisations,* whether public or private, collect, use and disclose personal information, the assumption being that a hierarchy exists. In a *distributed* system that relies on a membership permissions and validation process that assumption breaks down.

Let me be clear – I am not a critic of these technologies and do not discount their ability to deliver privacy benefits. I am simply pointing out that we need to understand more about them before declaring them the answer to our privacy dreams.

Consent technologies

Finally, another body of work is exploring ways in which technology can be used to underpin the core privacy concept of consent and the collection, use and disclosure of personal information for the purpose for which it was collected. This body of work relies on applying forms of digital rights management – a technology that fell into disrepute after the way it was used by the entertainment industry to prop up its decaying business model – to personal information. It also can, in some implementations, use semantic web principles to the same end. Essentially these approaches attach permissions to personal information and enable automated negotiations between information subjects and information recipients about the collection and subsequent use and disclosure of the subjects' personal information.

This is the type of approach that Jo Cooper and Cloud Insurance are pursuing through consent receipts aggregation technology. They are in distinguished company: this type of approach has been advocated by Professor Alex Pentland of MIT who has supported putting 'the individual much more in charge of data that's about them. This is a major step in making Big Data safer and more transparent, as well as more liquid and available, because people can now *choose* to share data.'[8]

Pentland also believes that personal information can be owned by data subjects, that there should be a proprietary right in personal information. This is a view that is expressed fairly often in US privacy discourse but not elsewhere. I think it's an interesting idea but that there is no chance of it becoming a reality. That said, the idea of giving individuals technological tools to negotiate personal information transactions needs to be explored and considered carefully.

The challenge for this audience is to understand and scrutinize the technologies and their implications realistically and from multiple viewpoints. As the recent ABS debacle has shown, Australians care very much about the privacy of their personal information and are unlikely to trust solutions that do not strike the right balance between functionality and protecting their rights.

They are also skeptical of government. Despite the Commonwealth Minister responsible for the census, Michael McCormack, commenting that the census was just like Facebook and dismissing concerns about the census enabling government to track the population as being 'much ado about nothing'[9] a significant part of the population thought otherwise. The same argument was used in the data retention debate when the (then) head of ASIO said:

> Are you arguing that it is OK for Microsoft or Google to profile you in order to sell you a new BMW, or some beauty product, that is alright for them, but it's not alright for the government on a very selective basis to access telecommunications metadata in order to save lives? That to me is a very

[8] See https://hbr.org/2012/10/big-datas-biggest-obstacles

[9] See http://www.smh.com.au/federal-politics/political-news/minister-says-census-no-worse-than-facebook-as-nick-xenophon-risks-jail-20160808-gqnobg.html

> distorted and worrying argument.[10]

There is a body of opinion in Australian governments to the effect that 'if you let the private sector do it, then government should be able to as well.' Frequently, this is the opinion expressed by those responsible for 'innovation' or 'disruption' agendas, and who see no boundaries to government information sharing.

But governments are different. Often we deal with them only because we are forced to do so. They can arrest you and they can punish you. The most frequent 'clients' of government are the most vulnerable and marginalised: they have no choice and, unlike Facebook, are not provided with any technology controls or settings. Our skepticism grows in proportion to the claims about personal information that are made by government that prove to be inaccurate, such as the 'opt-in' promise for the Personally Controlled Electronic Health Record or claims that the census had the best security features.

The path ahead

My approach to the tasks entrusted to me by the Special Rapporteur is of open-minded skepticism: the focus will be on evidence-based critical analysis. Mine is also an inclusive and collaborative approach – I won't succeed without the support of people like you.

While I am open-mindedly sceptical, I am also cautiously optimistic. I've learnt the trade-off arguments pervade the data protection discourse. I've learnt that they are almost always false, like the supposed trade-off between privacy and security. I don't think that there is room for an argument that says 'you can have big data or data protection, but not both.'

Finally, I'm also interested in developing a document that feeds in to the other

[10] See http://www.ethics.org.au/on-ethics/blog/march/exclusive-ex-head-of-asio,-david-irvine,-on-data-r

themes identified by the SRP. I’m particularly interested in how this work can contribute to questions about international regulatory norms and instruments as the preferred way to develop an international regulatory framework and to provide a greater amount of certainty to those who are developing the technologies that are designed to bridge the gap between big data and open data benefits and risk mitigation strategies.

David Watts
Commissioner for Privacy and Data Protection (Victoria)
7 September 2016