Who's afraid of the bad, big data? You might want to read this

May 28 2013, by Rob Livingstone

Shhh – there's so much buzz around your digital footprint. Credit: Yael P

Privacy and technology go together like music and dance: it's only when both work well together that the magic happens. But what about privacy in the age of big data, an era in which your every move has been recorded somewhere in the digital world through your electronic transactions?

Does the fact we're churning out ever greater volumes of data mean we are safe, by virtue of pack anonymity, or are we at risk of serious violations of the individual's privacy rights?

Your personal digital footprint – that indelible record of your every interaction in the electronic world – is just a tiny drop in the ever increasing sea of global data.

Your email traffic, internet search history, the geotagged images you take on your smartphone and share through social media sites, your retail purchases, loyalty program transactions, payments, road toll payments and medical records – to name but a few – are all part of the unique tread that makes up that footprint.

In addition to personally generated data, business systems' transactions, social media sites, healthcare, research, government and scientific agencies, together with myriad other sources, generate rivers of data that eventually flow into the various data centres dotted around the globe.

Many are owned by internet giants such as Google, Microsoft and Amazon, with others owned by multinational corporations and governments.

And that's just the start. Consider for a moment just one of the latest major scientific endeavours, the Square Kilometre Array (SKA) Telescope – a state-of-the-art radio telescope currently in development in Australia and South Africa.

Once operational, the SKA is expected to produce data equivalent to between ten and one hundred times the traffic of the entire internet, and will require the equivalent processing power of about a hundred million PCs! That's the shape of things to come …

Flicking the Vs

Simply put, big data is data that's too large or complex to be effectively handled by standard database technologies currently found in most

organisations.

For data to be regarded as "big", it should possess three key attributes – volume, velocity and variety:

- Volume is just what it sounds like: lots of data. To put this in context, YouTube users upload 48 hours of new video every minute of every day.
- Velocity occurs where the data is time-sensitive and needs to be processed and stored quickly. One example is the real-time profiling of internet display adverts that are customised according to your usage pattern.
- Variety covers the various forms that data can take, from neatly-structured tabular data, to unstructured data containing items such as images, emails, spreadsheets, social media conversations and streaming media. Currently, there is no universally accepted "one-size-fits-all" approach to handling this data variety.

Converting coal to diamonds

Data alone is of limited value. Only when the disparate array of data sources is merged, consolidated, analysed and interpreted does its potential value emerge.

For instance, climate records, geographic, population census data, medical records and other data sources could be used to identify and predict the trends of specific diseases.

Credit: Ownipics

For individuals, this has the potential to significantly improve access to more relevant, timely and accurate information stitched together from a range of sources.

In commerce, big data has the potential to translate into big value. In an earlier article on The Conversation I suggested our digital footprints have realisable value to others – in other words, that you are essentially the product, not the things you buy.

Big data and privacy

To comply with relevant privacy legislation, data that is to be externally released for purposes such as marketing, analysis and reporting should have the individual's personal information removed – a process known as anonymising, or deidentifying.

But when disparate data from a range of anonymised, independent data sources can be matched using specialised algorithms to geotagged information, it may be possible to reidentify data that was previously anonymised.

A number of researchers have already shown reidentification to be possible by using specially crafted matching algorithms.

The risks associated with the possible reidentification of personal information should be a topic high on the agenda for industry regulators, legislators and those concerned about information security and privacy.

On the bright side, the reidentification of big data is a distinct advantage for anti-terrorism and law enforcement agencies. The ability to pinpoint individuals who are a likely threat to society or involved in criminal activities would be largely seen as a positive use of big data.

But the possibility of misidentification is real, which may have serious consequences for the individuals concerned. Factors such as the provenance and accuracy of source data, together with the validity of the analytical techniques used, needs to be meticulously verified to minimise the occurrence of misidentification in such instances.

Laws of the land

Using overseas cloud computing providers such as Google or Microsoft to store and manage large data sets introduces the additional complexity of international data residency legislation.

The laws of the country in which the data is located (as opposed to where it is generated) apply. Managing the array of privacy and related data-residency legislation across multiple international legal jurisdictions is not a trivial exercise.

Data governance is a crucial consideration. The process of managing, analysing and interpreting big data involves the merging of the disparate data sources to newly created, consolidated data sets.

A number of questions arise over the governance and security of these new data sets, such as: who owns, or has title to, these newly created, aggregated data sets?

Who decides what access controls should be applied to the new data sets, and in which legal jurisdiction?

Globally, cybercrime is a multi-billion dollar business with some of the smartest brains employed to crack security systems. Put simply, there is an ongoing arms race between the cloud providers and the cybercriminals, and sometimes the latter win.

We should never ignore the fact big data presents a rich target of opportunity for cybercriminals.

Big data is currently the "hot topic" in the IT world, and as is the case with all preceding technology innovations and fads, has whipped up its fair share of emotions.

The concept of the technology hype cycle – coined by the American IT

firm Gartner, Inc in the 1990s – is frequently used to describe and map the typical phases in the evolution and adoption of new technologies.

The five-phase cycle begins with a "technology trigger" (the breakthrough of a product or technology) and ends with a "plateau of productivity" (the point at which the technology or concept becomes an accepted and stable part of the landscape).

Between those bookends are a couple of stages with self-explanatory titles: the "peak of inflated expectations" and the "trough of disillusionment".

Big data, some argue, may have already passed its peak of expectations and be heading for the trough.

Business and government agencies are, understandably, looking to maximise the potential value inherent in big data; but while mitigating the numerous data integrity, privacy and security risks, they should not be deafened by the current buzz around big data.

The question when that buzzing stops is whether you will be better off in the volatile, globalised world of big data.

This story is published courtesy of The Conversation *(under Creative Commons-Attribution/No derivatives).*

Source: The Conversation