Using ‘Big’ Metadata for Criminal Intelligence: Understanding Limitations and Appropriate Safeguards

Alana Maurushat
Faculty of Law, UNSW
Sydney, Australia
+61 2 9385 8027
a.maurushat@unsw.edu.au

Lyria Bennett-Moses
Faculty of Law, UNSW
Sydney, Australia
+61 2 9385 2254
lyria@unsw.edu.au

David Vaile
Faculty of Law, UNSW
Sydney, Australia
+61 2 9385 3589
d.vaile@unsw.edu.au

ABSTRACT

Using Internet Service Provider ‘Big’ metadata as a case study, we examine legal and ethical issues with machine learning Big Data tools developed and deployed in Australia for law enforcement intelligence purposes. In order to do this, we outline the benefits, limitations and risks of these tools, analyze current methods for de-identification and anonymisation, and discuss necessary safeguards.

Categories and Subject Descriptors

Open data, and big data

Keywords

Big Data, Open Data, Metadata, Communications Metadata, Privacy, Machine Learning, Law Enforcement, Criminal Intelligence,

1. INTRODUCTION

Government agencies are developing sophisticated technologies including search algorithms and Big Data analytics software that will allow metadata and other information to be collected and analyzed [10]. Data sets are governed by regulatory obligations around what data is collected, who is permitted to access and use the data, other security and privacy obligations. Other data is generated by one government unit and used exclusively by the organisations. As such, data available to one government agency may not be available to another agency even if the data is critical for a particular criminal intelligence task. There are many reasons why data is not freely shared between agencies, most notably concerns to protect privacy, communications confidentiality and personal information security, to minimize state encroachment into the private life of its citizens, to avoid discrimination, and to prevent misuse.

Big Data tools have potential to alleviate some of the concerns around unauthorized practices in data collection, use, re-use and analysis as such tools allow for de-identification and/or anonymisation. For example, data can be de-identified and anonymized to reduce concerns around the collection and use of personal information, while other Big Data tools allow for automation that removes the need for humans to see the data (including personal information). This paper will address legal and ethical issues with the use ‘Big” metadata in criminal intelligence. We will start by providing an overview of a theoretical case study of machine learning algorithms running on Australian Internet Service Provider metadata under data retention policy. We then examine the current legal framework for Big Data tools using metadata stored by ISPs and address the possibility of allowing machine learning tools to analyze metadata for criminal intelligence purposes. We address privacy issues by looking at de-identification and re-identification, noting where AI techniques could assist in minimizing privacy risks.

2. METADATA FRAMEWORK

“Metadata” is referred to as “telecommunications data” in Australia. The term is currently undefined but is generally considered to be information or documents that are neither ‘stored communication’ nor ‘content’, meaning most information that does not embody the substance or content of a communication such as an email, phone call or use of an online service .

The Minister for Communications tabled the *Telecommunications (Interception and Access) Amendment Bill 2014* (the Data Retention Bill) on October 30, 2014. This Bill would establish mandatory two year data retention by carriers and ISPs of all users’ metadata.

“Telecommunications Data” is said to *exclude* search browsing and the content of communications [2]. Currently there is no constitutionally enshrined right to privacy in Australia, nor any relevant legal right such as the tort of invasion of privacy [18]. Freedom of speech and communications confidentiality similarly lack protection [18]. Instead a variety of laws, including telecommunications-related statutes and procedural aspects of state and federal crime statutes, create state obligations around due process and reasonable search and seizure. Law enforcement must comply with these obligations. They are, however, largely excluded from the data protection obligations in the *Privacy Act 1988,* as explained below.

ICAIL '15, Jun 08-12, 2015, San Diego, CA, USA
ACM 978-1-4503-3522-5/15/06
http://dx.doi.org/10.1145/2746090.2746110

Australia protects personal information in the form of Privacy Principles. These principles are: Open and transparent management of personal information; Anonymity and pseudonymity; Collection of solicited personal information; Dealing with unsolicited personal information; Notification of the collection of personal information, Use or disclosure of personal information; Direct marketing; Cross-border disclosure of personal information; Adoption, use or disclosure of government related identifiers; Quality of personal information; Security of personal information; Access to personal information; and Correction of personal information .

"Personal information" means information or an opinion about an identified individual, or an individual who is reasonably identifiable: (a) whether the information or opinion is true or not; and (b) whether the information or opinion is recorded in a material form or not [39]. Australian courts have yet to rule whether metadata is personal information under the *Privacy Act,* although in the Bill, the aim seems clearly to use metadata for the purposes of identification.

While Australian law enforcement and intelligence agencies are generally not subject to the obligations in the *Privacy Act*, let alone a constitutional or legally enforceable right to privacy, this does not mean that privacy and related concerns are unimportant, or should be disregarded. The Bill has been critiqued as not being sufficiently narrow in its focus, making it unreasonable and disproportionate despite the 39 amendments inserted by the opposition Labor party. For example, there is no warrant required to access metadata, and there is no requirement to delete data once investigation is complete. Other comments have queried the evidence for effectiveness of retained metadata for the initially stated purposes centered on anti-terrorist activity. Many have asked that the bill be circumscribed to ensure that privacy is protected and the retention and use of communications metadata is proportionate [1]. Currently, the bill neither provides technical or legal safeguards, nor requires any review of its efficacy, or the efficacy of the scheme it establishes.

The Data Retention Bill was passed by the Australian Parliament in March 2015 which has opened the door for the employment of data analytic tools on high volume communications metadata.

3. USES AND LIMITATION

Data could be used in three general ways for law enforcement and intelligence. First and most common, Big Data is used in a specified and targeted investigation where an offence has been committed, the main suspect is known and information is sought about his/her communications. Second, Big Data *could* be used to help identify unknown suspects after an offence has been committed. In this instance the perpetrator of the crime would be unknown but patterns produced from Big Data tools *could* aid in identification. In both the first and second instances, the identity of the person attached to the data is critical. The third use, and the focus of this paper, is the use of Big Data to feed into analytic tools where information patterns emerge but where there is no identified or unidentified suspect prior to deployment of the tools; in some cases there has been no offence committed. In this instance, Big Data tools would use vast amounts of unconnected and disparate data, and apply machine learning pattern matching and other techniques to discern patterns in it and propose associations. Attempting to predict patterns of behavior, however, is both the most extreme promise of Big Data (in the form of "predictive analytics"), and its most potentially controversial over-reach.

While predictive pattern analysis claims to predict the future, it can only do so probabilistically based on analysis of historic data. Historic data is used to train machine learning algorithms in order to classify future instances or identify trends. The process assumes stability in relation to either the correlations themselves or their rate of evolution (so that anticipated changes over time can be built into the model). Such an assumption is necessary for information derived from historic data to be used to draw conclusions about the future.

In the case of communications metadata, social network analysis can be used to determine relationships between individuals and identify central nodes. As well as being used for surveillance and identification of suspects and their networks, this can also be used (with other data) for the third purpose of identifying general patterns. For example, different communication patterns may correlate with particular forms of criminal activity. In this case, one is using historical correlation for prediction, and assuming stability in patterns over time [16].

Even where there are good reasons to assume stability in a particular context, information about correlation that lacks an understanding of causality cannot predict the impact of intervention in the system. This is best illustrated in Pearl's discussion of causality, which identifies a "cause" with the ability to change an effect variable by surgically altering the cause variable [21]. The ability to predict the likelihood of a profiled individual committing a terrorist act or other offence based on communication patterns may shift as the existence, strategy and rhetoric of different terrorist groups evolve, as communications technologies change, and as events happen in the person's life or mind [6]. Not all of these variables will be captured in the data.

In predictive analysis, feedback can occur where an intervention generates negative consequences for individuals engaging in behaviors that are correlated (or thought to be correlated) with crime, leading to avoidance of those behaviors (without changing the likelihood of engaging in criminal conduct). This can occur where profiling is based on features that are susceptible to change (like membership of a club, hairstyle, clothing worn, places visited, language spoken in public, communication patterns).

Another feedback loop can occur when interventions lead to systematic bias in the dataset being analyzed [17]. For example, if there are changes in locations where police are deployed or surveillance targets (due to predictive policing) then crimes that take place in those locations or through the activities of those targets are more likely to become part of the dataset that is analyzed in future iterations of the algorithm. This makes it possible that "hot spot" locations identified will continue to be treated as "hot spots" despite *actual* changes in the locations where crimes are taking place.

The reliance on information about correlations derived from "black-boxed" algorithms also makes it difficult to know when a machine learning algorithm has "learned" to discriminate based on prohibited categories such as race. Even if one develops an algorithm to ignore data as to a person's race, other data collected (such as physical features, address, language spoken, communications patterns and so forth) may well correlate with race. Where any of these features correlates with a particular behavior or tendency, it will be "learnt" by a machine learning algorithm as a means of predicting that behavior or tendency. This will be so even if the feature concerned has no causal relationship with the behavior, due perhaps to the fact that the behaviors of particular groups or neighborhoods are more likely to come to the attention of police, and hence form part of the analyzed data.

While none of this implies that data analytics should be abandoned as an enterprise, it does suggest limitations in its utility. Such limitations must be taken into account in weighing the benefits of police access to large volumes of data against the privacy implications of allowing such access. It also suggests a need for caution in decision-making that relies on inferences drawn from data analytics, particularly where there are implications for civil liberties, including the risk of stigmatizing discrimination against minority groups.

4. DE-IDENTIFICATION, ANONYMISING AND RISK OF RE-IDENTIFICATION

There are some safeguards which could work to reduce privacy concerns and discriminatory profiling. De-identification and anonymizing data is a limited safeguard. For our purpose, "de-identification" is the removal, stripping or obfuscation of directly identifying elements from a data record or set such that the result is not immediately identifiable as associated or linked with a particular individual.

"Anonymization" is the process of rendering data into a form that does not identify individuals in circumstances where identification is not likely to take place.

The main difference between de-identification and anonymisation is that de-identified data may potentially be re-identified, whereas anonymisation is said to be irreversible [6]. A de-identification technique allows for re-identification, while the law sets the parameters and context when data can be re-identified. With anonymization, the technology and the regulatory framework forbid re-identification (this may not always prove successful but the goal is to not allow re-identification under any circumstance).

"Re-identification" is the ascertainment or creation of a link between a particular individual and a set of data in circumstances which had initially, either deliberately (after de-identification) or by chance, appeared not to support making any such link.

"K-Anonymity" is a popular approach for data anonymization, and has been used for health information data [21]. K-anonymity is the anonymization of a record whereby the probability of the record being re-identified is *1 / k* [5] [13] [14] . K-anonymization has been explained like this:

"A k-anonymized data set has the property that each record is similar to at least another k-1 other records on the potentially identifying variables. For example, if *k = 5* and the potentially identifying variables are age and gender, then a k-anonymized data set has at least 5 records for each value combination of age and gender" [14].

K-anonymization uses a number of techniques to transform data including suppression, randomization, irreversible coding, reversible coding, generalization, and tagging [13].

The term K-anonymization is somewhat deceptive in that it can be used as both a de-identification tool and an anonymization tool depending on the methods and value of K used in the process. Typically, anonymizing methods would create a lesser risk of re-identification than would a de-identification tool. Regardless of which technique is used K-anonymization data custodians would then examine a data set, and select a value of *k* that is commensurate with a risk of re-identification appropriate to the gain. This risk is often referred to as "threshold risk" or "threshold risk metrics" where re-identification and privacy intrusion are part of the risk.

If one assumes that privacy is paramount and re-identification risk is to be as low as possible, then the selection of a *k* variable to minimize re-identification is a sensible choice. The problem is that effective de-identification can lead to information loss, arising from the deliberately introduced distortions to the data [14].

In the same way that there is often said to be an inescapable trade-off between IT security and user convenience, there is also potentially a trade-off between privacy (in the form of de-identification) and data richness. The benefit of privacy is for one stakeholder (the data subject), while the benefit of the rich data is for another (the analyst, or custodian, if it's in house). It may be a fundamental conflict. This suggests that differing stakeholders may prefer differing levels of de-identification, depending on how they weight the potential loss of information richness against the potential harm from re-identification.

While K-anonymisation and differential privacy are the leading techniques at present, other anonymization techniques such as substitution, shuffling, number variance, data variance, character masking and cryptographic techniques are also commonly used [22].

All of these techniques, however, still suffer from the risk of re-identification. The now proven reality is that de-identification and anonymization simply do not work where significant amounts of data (high-dimensional data) are involved [11] [15]. Notably, high dimensional data sets such as ISP metadata are of interest to law enforcement and intelligence making re-identification highly probable.

One recent promising yet not yet conclusive avenue is 'differential privacy'. The privacy risk in differential privacy is quantified by the following formula [12] [16] [23]:

A randomized function K gives ε-differential privacy if for all data sets D and D′ differing on at most one row, and all S ⊆ Range(K),

$$Pr[K(D)\ S] \leq exp\ (\mathcal{E})\ X\ Pr[K(D')\ \mathcal{E}\ S]$$

The above equation for differential privacy is the equivalent of "personal information in a large database that is not modified or released. Instead, a third party, such as a researcher, can submit questions about the information in the database by going through an intermediary piece of software that serves as a privacy guard" [7]. If the threshold risk is too high, the researcher may choose to abandon the query, or to modify the algorithm to further reduce privacy risk through techniques such as noise where, for example, additional records are added to the system thereby reducing re-identification risk [12] [13] [14] [23].

Tockar applied differential privacy theory by using some photos and insights combined with the open TLC taxi dataset, and was able to re-identify 11 individuals [22].

Tockar's methodology was criticized as insignificant given that his sample size was extremely small, considering that there were 173 million taxi rides and only 11 individuals were identified. As Daniel Barth-Jones put it, "when does 99.99999936% equal zero?" [4]. Barth-Jones further argues that the taxi data had been shown to have used insufficient de-identification means and that encrypted methods would have greatly reduced re-identification.

Both Tockar and Barth-Jones show deficiencies in their approaches [5]; Tockar did not publish the sample size and Barth-Jones' 11/173 million is equally problematic as we can assume that Tockar's sample size was not 173 million taxi rides.

Whatever de-identification or anonymizing technology is used, the fundamental question remains the same: what is the threshold risk of re-identification and how to we reduce risk?

As metadata offers rich personal information for analysis, privacy concerns around protection of personal information and due process can be met with law enforcement and intelligence use of metadata.

5. LIMITING RE-IDENTIFICATION

Limiting re-identification alleviates some of the concerns around unreasonable search, due process and privacy. Limiting re-identification where data has been de-identified or anonymized can be done through technical means (e.g. use of differential privacy) and/or by regulatory means. Regulations could be passed providing clear guidance when re-identification is permitted and under what circumstances.

The New Zealand Privacy Commissioner recently proposed a specific prohibition on re-identification

of de-identified data: that it be a criminal offence to avoid the efforts made to protect what was once personal information from later unwanted de-identification.

Given the importance that effective de-identification plays for safe publication of data, a provision of general application should be enacted to implement this model.

An alternative to criminal prohibition is to provide other legal remedies in the event of a negative outcome from re-identification. Other possibilities look to mandatory data breach disclosure notification, cause of action for serious intrusion into privacy [3], and ethical codes of conduct [8]. Ethical codes are more difficult to implement than legal remedies as measuring ethical metrics for compliance is convoluted [9].

6. PITS AND PETS

There's no question that many Big Data tools will be privacy-invasive technologies (PITs). Big data tools are arguably the most privacy invasive technologies ever developed if used without restraint. The hope, however, is that Big Data tools could also be used as a safeguard against disproportionate privacy concerns. Privacy enhancing technologies (PETs) could be developed and embedded into the tools [24]. The obvious example is anonymysing and encrypting metadata making re-identification more difficult. Software agents could be layered into the tools to enhance privacy according to the type of data, and combination of use [24]. As the probability of re-identification becomes more likely, the underlying algorithm could signal that different methods of protection are required. This is an area that is ripe of re-thinking privacy and the development of PETs.

7. MACHINE LEARNING

Many Big Data tools have been developed for law enforcement, with different nations opting for different controls around both the data and decisions made based on the data. For example, France applies strict control on Big Data legal tools requiring personal control over the data and decisions made. Spain is more lenient allowing automated machine learning tools to analyze the data as well as make decisions based on the data [8]. There are no standards in Australia for this type of activity.

One advantage of insisting on machine learning is that in theory no human eyes would need to see the data. This potentially reduces claims of intrusive privacy invasion, and individual abuse around collection, processing, storage, use and re-use. Automated machine learning coupled with limitations on re-identification can reduce abuse around personal information.

8. CONCLUDING REMARKS

Big Data tools promises improved policing and intelligence. This leads to calls for the capture, retention and analysis of new datasets, such as the drive to enable law enforcement access to communications metadata in Australia. The technologies and tools available will grow quickly in this field with both benefits and risks. This article gave a layout of some of the safeguards that will be required.

9. REFERENCES

[1] 15th Report of the 44th Parliament, Parliamentary Joint Committee on Human Rights (October 2014) http://www.aph.gov.au/~/media/Committees/Senate/committee/humanrights_ctte/reports/2014/15_44/15th%20Report.pdf

[2] Attorney Generals Department, "Data Retention Bill – Proposed Data Set", October, 2014, at http://www.attorneygeneral.gov.au/Mediareleases/Documents/DataRetentionDraftDataSet30October2014.pdf

[3] Australia Law Reform Commission, "Report into Serious Invasions of Privacy in the Digital Era", September 2014, at http://www.alrc.gov.au/publications/serious-invasionsprivacy-digital-era-alrc-report-123

[4] D. Barth-Jones, "The Antidote for 'Anecdata': A Little Science Can Separate Data Privacy Facts from Folklore," Info/law Blog, Harvard University, November 21, 2014, at http://blogs.law.harvard.edu/infolaw/2014/11/21/the-antidote-for-anecdata-a-little-science-can-separate-data-privacy-facts-from-folklore/

[5] R. Bayardo, and R. Agrawal, "Data Privacy through Optimal K-Anonymization", Proceedings of the 21st International Conference on Data Engineering, 2005athttp://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1410124&tag=1

[6] J. Berman, *Principles of Big Data: Preparing, Sharing and Analyzing Complex Information*, Elsevier, 2013, p 229.

[7] A. Cavoukian, and J. Jonas, *Privacy by Design in the Age of Big Data*, June 8, 2012, at http://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf

[8] P. Casanova, "CAPER Regulatory Model: Platform to Fight Organised Crime" AustLII Workshop on Privacy, Sydney, September, 2014, at http://www.austlii.edu.au/austlii/seminars/2014/3.html

[9] C. Connolly, and D. Vaile, "Drowning in Codes: An Analysis of Codes of Conduct Applying to Online Activity in Australia," Cyberspace Law and Policy Centre, March 2012, at http://cyberlawcentre.org/onlinecodes/report.pdf

[10] T. Davies, "Open Data Policies and Procedures: An International Comparison", October 23, 2014, at http://dx.doi.org/10.2139/ssrn.2492520

[11] De Montjoye, Yves-Alexandre,"Unique in the Crowd: The Privacy Bounds of Human Mobility" *Scientific Reports* 3, 2013, at

http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

[12] C. Dwork, “Differential Privacy” 33rd International Colloquium on Automata Languages and Programming (ICALP), Lectures in Computer Science, 2006, at http://dx.doi.org/10.1007/11787006_1

[13] K. El Eman, *Guide to the De-Identification of Personal Health Information*, CRC Press, 2013, at. http://www.crcpress.com/product/isbn/9781466579064 (by subscription)

[14] K. El Eman, and F.K. Dankar, “Protecting Privacy Using K-Anonymity” *Journal of the American Medical Informatics Association* Vol 15 Issue 5, 2008, at, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029/

[15] E. Felten, and A. Narayanan, “No silver bullet: De-identification still doesn’t work” July 9, 2014, at thttp://randomwalker.info/publications/no-silver-bullet-de-identification.pdf

[16] B. Harcourt, *Against Prediction: Profiling, Policing and Punishing in an Actuarial Age*, University of Chicago Press, 2007, at http://www.bernardharcourt.com/documents/readers-companion.pdf

[17] J. Lee, and C. Clifton, “How Much is Enough? Choosing Ɛ for Differential Privacy,” Lecture Notes in Computer Science, Springer-Verlag Berlin, Vol 7001, 2011, at , https://gnunet.org/sites/default/files/Choosing-%CE%B5-2011Lee.pdf

[18] A. Maurushat, “Australia” in Cooke, S. (ed) *Internet Freedom* (Freedom House 2014).

[19] National Statistical Service (NSS) Confidentiality Information Series. http://www.nss.gov.au/nss/home.NSF/pages/Confidentiality+Information+Sheets

[20] B. Raghunathan, *The Complete Book of Data Anonymization: From Planning to Implementation*, CRC Press, 2013, at http://www.crcpress.com/product/isbn/9781439877302 (

[21] J. Pearl, *Causaltity: Models, Reasoning, and Inference*, Cambridge University Press, 2009, at http://bayes.cs.ucla.edu/BOOK-2K/

[22] A. Tockar, “Tiding with the Stars: Passenger Privacy in the NYC Taxicab Dataset,” Neustar Research, September 15, 2014, at http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/

[23] A. Tockar, “Differential Privacy: The Basics” Neustar Research, September 8, 2014, at http://research.neustar.biz/2014/09/08/differential-privacy-the-basics/

[24] U.S. Public Policy Council of the Association for Computing Machinery, “Big Data and Consumer Privacy in the Internet Economy” 79 FR 32174, 2014.