Center for Cyber Security and International Relations Studies

Bulk Data: Intelligence and Surveillance

Marcelo Malagutti[1]

Abstract

Up to the 1990's, surveillance was an expensive and challenging task, making it necessarily contained and focused on a limited number of 'targets'. The cyber realm and the contemporary digitisation of human everyday life activities, the increasing computing and storage capacities, and their declining costs, made it easier and more effective.

Snowden revelations showed to the world that intelligence services around the globe were systematically collecting, processing, and even exchanging, bulk data regarding personal communications, in most cases without legal warrants. The following debate left clear the growing power of private data collection for surveillance purposes, and the importance attributed to these activities by the security and defence establishments.

The struggle between the need for security is often presented as opposed to the need for liberty, and this has always been the central point of the debate about mass surveillance.

This article summarises and clarifies many of the involved concepts, and translates some of the technical foundations into a non-technical (or less-technical) language, and presenting the evolution of mass surveillance and the current situation of the political debate regarding it.

Key-words: Bulk Collection; Metadata; Mass Surveillance; Signals Intelligence;

[1] Ph.D. Candidate at Brazilian Army War College (ECEME)

Introduction

In 2013, Edward Snowden, a subcontractor's employee of National Security Agency (NSA), revealed that the agency and its Five Eyes partners (Australia, Canada, New Zealand and the UK) were, allegedly, running programs of mass surveillance (Greenwald, 2014). In the hype that followed it has become clear that Five Eyes indeed had processes for bulk data and metadata collection and analysis. Still today there is an active debate focused on the importance of maintaining these bulk powers and on its actual value for intelligence gathering.

This article shows that bulk data interception, collection and analysis is useful for intelligence gathering, although having limitations. The first section provides a very brief history and definition of signals intelligence. The second presents a few basic concepts related to modern data communication protocols. A third one conceptualises metadata. The fourth section explains bulk data collection and analysis. Then the limitations of bulk collection and its utility are discussed. The seventh part discusses some legal battles regarding bulk collection and civil liberties. Finally, some conclusions are drawn, showing that bulk powers are indeed relevant for security purposes.

A Very Brief History of Signals Intelligence

Communications Intelligence (often referred to as COMINT) has always played a significant role in security and defence matters. To avoid his enemies' interpretation of possibly captured messages, the Roman emperor Julius Caesar (100 BC to 44 BC) is appointed as having already used a transposition cypher algorithm (Singh, 2000:pp.14-20).

A particular function of COMINT is Signals Intelligence (SIGINT). The British Signals Intelligence Agency GCHQ defines it as 'intelligence derived from intercepted signals' (GCHQ, 2011:p.9). SIGINT has become more and more relevant since the advent of the telegraph. GCHQ's website remarks the importance of the interception of the famous Zimmerman Telegram as one of the main reasons for the U.S. entering the first world war (GCHQ, 2016). Radio communications made SIGINT even more important, and GCHQ's website also points the history of Bletchley Park, where Alan Turing and his team created Colossus, the first computer in history, which helped to decipher the German Enigma code, an important asset for winning WWII. The entire operational structure of Bletchley Park was based on 'Passive SIGINT', with the interception and transcription of every radio message sent by the Germans (bulk interception or collection), for subsequent analysis.

The process nowadays is quite similar (GCHQ, 2011:pp.9-12). In recent years, 'the Internet is a major source of comparable intelligence power today' (Omand, 2015). For the NSA it has become even easier:

> As the Internet developed, a large portion of the Internet backbone passed through the United States, meaning that many foreign-foreign communications could be accessed by surveillance done inside the US. Previously, foreign-foreign communications would have been accessed outside of the US, where the US Constitution and various laws are less strict than for access inside the US (Swire, 2015).

Communication Protocols

A bit of technical background regarding data communication is necessary before we proceed. The analogy, as usual, is based on the traditional mail letter. There is an *envelope*, on which outside is the protocol information necessary for the postal service to deliver the mail; usually the name and address of the receiver, in the front, and the name and address of the sender in the back. What goes inside the envelope, the *message*, the actual data, or *payload*, is not (or in general should not be) viewed by the postal service. Thus, a protocol defines the 'syntax' and the 'semantics' of the data used to establish the communication but does not affect or work with the payload.

The Open Systems Interconnection (OSI) model specifies seven different layers of networking interconnection: (i) physical; (ii) data link; (iii) network; (iv) transport; (v) session; (vi) presentation; (vii) application (ISO, 1998). The first layers comprehend the most basic features of communication devices, while the latter ones are closer to the final user. For each layer, there are several different protocols available. Let us consider a few popular examples present in almost every smartphone nowadays. Global System for Mobile Communications (GSM) [2] is the protocol majorly used by the mobile phone networks to implement its services over the two first layers, while the Internet Protocol (IP) is implemented as the 3rd layer, and the Transmission Control Protocol (TCP) as the 4th one. The HyperText Transmission Protocol (HTTP) that carries HyperText Markup Language (HTML), used for web browsing, and the Extensible Messaging and Presence Protocol (XMPP), used by some messaging applications, are implemented as the 7th layer. The ISO8583 protocol used for credit or debit card transactions processed on a wireless Point of Sale (POS) terminal is also at the 7th layer.

Each superior layer is implemented on the *payload* of its supporting layer. Thus, the HTTP protocol (both *envelope* and *payload*) is the *payload* of the data packets of the TCP protocol. The latter implements its *envelope* and *payload* on the *payload* of the IP data packets *payload*, and so on. For every packet of data in a higher layer there is at least one packet in the lower layers, but usually more.

As a rule of thumb, protocols work by establishing a connection, sending and receiving data, and closing that connection. Establishing a connection is often a resource and time-consuming process. The client asks for it sending a message (a data packet), the server allocates its resources to deal with the connection, sends its 'synchronisation acknowledge' message (another data packet) to the client, and this responds with its 'acknowledge' message (a third data packet) to the server. This process is also known as *handshaking*. Closing a connection is another 'expensive' process (regarding computational resources and time), and can involve four protocol data packets.

To avoid this cost of *handshaking* and *connection closing*, many communication systems try to maintain the connection open as long as possible. A usual technique consists of using *keepalive* messages from time to time: a message with the only purpose of generating traffic, so the server does not disconnect it. Even so, once the connection is established, it can be lost for different reasons. For instance, after an extended period of inactivity, the server can decide to close it for freeing resources to be used elsewhere. The connection might also be interrupted by 'lack of signal'. Then, when trying to send a *keepalive* message or a data message, a protocol 'connection lost' error will be generated, and the application will try to re-establish it with a new *handshake*, often with the user not even noticing it.

[2] Here GSM generically refers to the infrastructure that encompasses the original 2G mobile phones, the 3G Universal Mobile Telecommunications System (UMTS) and the 4G Long Term Evolution (LTE) protocols.

In the multi-layered communications structure, each protocol establishes its connections sending messages as packets of the inferior layers. Thus, in our Smartphone example, the web browsing HTTP packets generate TCP packets, which generate IP packets, which generate GSM packets.

Metadata

The classical definition of metadata is 'a set of data that describes and gives information about other data'[3].

Every protocol involved in the communication has its own metadata. For each protocol, metadata is the data that goes in the *envelope* part of the message. It is possible to identify, in the postal letter example, who sent a letter to whom. The name and address of both the sender and receiver are part of the metadata. The postal service, however, actually has three more information in its metadata: when (the *timestamp*) it has been posted and where, and when (another *timestamp*) it has been delivered. In a simple phone call, the same information is also available in its metadata. Additionally, how long the conversation has been. For every protocol involved in the communication, there is 'a bunch' of useful information available, even without looking inside the *payload* of that protocol's data packets.

A GSM Smartphone is basically a radio device that communicates with antennas in a cell, or base station, within a limited range. When it enters the area of a new antenna, it establishes a connection with it using its GSM protocol *handshake*. It identifies itself to the base station and thus provides information regarding its position at that time. This metadata is what allows the telecom company to account the minutes of talking, messages sent, Internet traffic and roaming usage of the smartphone, for billing purposes. It can also be used to allow investigations of where someone was at a particular time. Or with whom this someone has been talking or texting.

In GCHQ's terms, 'roughly, metadata comes from the part of the signal needed to set up the communication, and content is everything else' (GCHQ, 2011:p.10).

Bulk Data Collection and Analysis

The U.S. National Research Council has established a conceptual model for working with SIGINT as follows (NRC, 2015).

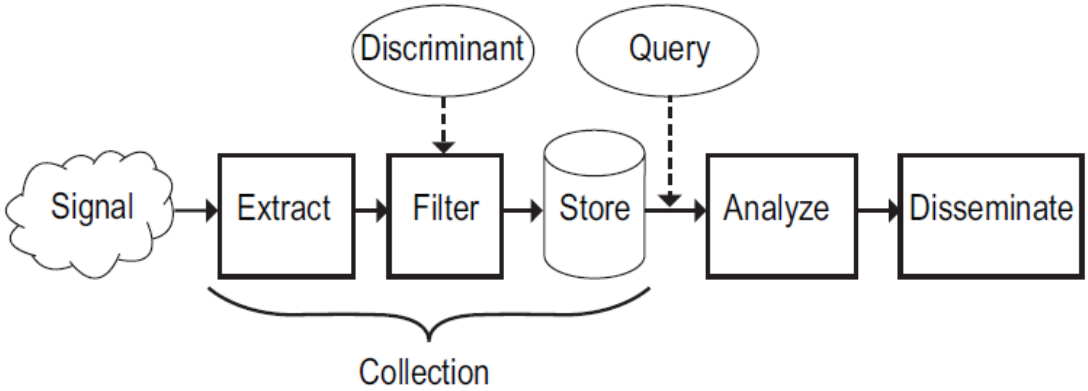

Figure 1 – U.S. National Research Council conceptual model for working with SIGINT

[3] http://www.oxforddictionaries.com/definition/english/metadata

The concept of *bulk data* varies according to the source, but the White House has defined it as the 'authorized collection of large quantities of signals intelligence data which, due to technical or operational considerations, is acquired without the use of discriminants' (U.S. White House, 2014). From Figure 1, where the *Discriminant* is applied in the *Filter* stage, *bulk data* can be both the *Signal* and the result of the *Extract* stage. The application of *discriminants*, thus, is what differentiates *bulk* from *targeted* collection.

Bulk data collection consists of the practice of gathering and recording everything that passes thru a network, for further analysis (as in the historic Bletchley Park's case). 'Bulk access can be achieved by intercepting terrestrial microwave links, satellite links and undersea cables' (Omand, 2015). However, if not interpreted, the bulk signals collected are meaningless, or only 'noise'. Only after somehow processing it data becomes useful.

A *sniffer* (or network analyser) is software that analyses network traffic *signal* and identifies the packages related to specific protocols, to *extract* and store them for further processing. The same bulk signal applied to different *sniffers* will provide different metadata, regarding the distinct protocols filtered. As a result of this process, there is a set of *bulk metadata* databases extracted, to which the further application of *discriminants*, for filtering the *target* information, can be applied.

Limitations of Bulk Collection

Rid (2015) has presented some possible limitations regarding the importance of bulk collection. One would be the increasing use of encryption in popular communications systems. Indeed, cryptography can be useful for protecting the content of messages. However, cryptography can only be applied to the *payload* area of each protocol; if the *envelope* area of the protocol is encrypted the network will not route it correctly, and then communication fails. Thus, encryption has a very limited effect, if any, on metadata.

Besides, peer-to-peer encryption of communications assumes that the content (the *payload*) sent will be precisely the same that is received. Thus, the *payload*, once encrypted, cannot be changed. So, even if the *payload* content cannot be inspected, it could be compared in terms of its *shape*: who has sent and received the *same payload* in its way from the sender up to the receiver.

One second limitation would be the rise of permanent connections for communications. As seen above, there are different levels of protocols, and when a GSM mobile enters the area of a distinct antenna, it establishes a new connection (a *handshake*) with this antenna creating GSM packets, and thus GSM metadata. It is possible that the connections of the protocols of the superior layers (IP, TCP or whatever) have not yet been lost, but, again, if the connection has been lost, it shall be re-established, probably in an automated way, often invisible to the user. But every handshake, *keepalive*, or data package sent, in every protocol, generates its metadata. The *keepalive* message used to maintain the connection open produces traffic of a particular nature in each and every protocol of the lower levels, creating its metadata. Thus, the fact that WhatsApp maintains its connection open for long periods of time does not mean it is not generating lots of metadata, with possible relevance for different publics.

The third possible limitation faced by bulk collection would be the rise of anonymity protection technologies as, for instance, Tor[4]. This author agrees that 'blocking such tools is technically difficult

[4] Tor is the acronym for The Onion Router, an open source anonymity network (https://www.torproject.org)

and politically unacceptable for a liberal democracy' (Rid, 2015). But *breaking* them is a different subject. As said above, it is possible to compare the *payload shape* (without opening it) in various communication events. With enough bulk metadata, comparing the *shape* of the payload, then identifying the route, and so the servers, and more importantly, where the message was generated and who was the final receiver, becomes feasible. This would not be possible with targeted surveillance if only the sender or the receiver were being 'watched'.

What is probably the most fundamental limitation for the use of bulk data must be the processing power needed to analyse and correlate all captured *bulk data*. The increasing connection of new devices, known as the Internet of Things, and the massive volume of generated applications data, named Big Data, pose enormous challenges for data analysis. Omand (2015) stated that 'the NSA touches about 1.6 percent of total Internet traffic, estimated at 1826 petabytes of information a day' of which 'only 0.025 percent is actually selected for review' and processing.

However, he also stated that 'the ability to cheaply transfer, store and mine digital data have all transformed the opportunities for obtaining secret intelligence.' Moore's Law, which predicts that computing power doubles at the same price, every two years, is still valid (Moore and Brock, 2006). Even those who predict that current computer design is reaching limits imposed by physics recognise that new technologies like quantum computing may provide significant increases in processing power (The Economist, 2016). Not less important, the development of new data mining and data analytics techniques allows intelligence services to deal with unprecedented amounts of data. Hadoop is the name of an open software platform inspired by Google's MapReduce (Dean and Ghemawat, 2004). It is designed to provide 'distributed processing of large data sets across clusters of computers using simple programming models'[5]. Not surprisingly, it is the platform GCHQ has chosen to keep its bulk events data. 'With hundreds of hard disks working simultaneously, multiple gigabytes can be read per second. This allows the processing of the multi-terabyte datasets we intercept' (GCHQ, 2011:p.60). The entire GCHQ data mining book, indeed, is focused on the research and development of new technical capabilities for dealing with massive bulk data.

A few examples of these tools can also be brought from Google, like those based on the hundreds of millions of search arguments submitted daily to its engine. Google Trends pinpoints the relative frequency of each argument, by region and language, and allows drilling down the trend. Google Insights has a similar concept, being used to offer hints for users, based on trends and their typing. Google AdWords offers advertising based on user recent searches. Google Correlate analyses the relation between search arguments and real-world trends. These are publicly available tools that exemplify possibilities in dealing with bulk data. Google BigQuery has been presented as 'a fast, economical and fully managed data warehouse for large-scale data analytics'. It is a commercial product that, among others, can be used to face Big Data processing issues.

NSA's new data centre in Utah, which cost 1.5 billion U.S. dollars, and operational since 2014, was built with the exact purpose of collecting and processing bulk data. Its web page in the NSA's Domestic Surveillance (DS) Directorate website does not specify its storage and processing capacity due to 'national security reasons'. During the site's construction, the press published a lot of discrepant news regarding the centre's storage capacity, from exabytes (10^{18} bytes) to zetabytes (10^{21} bytes) and even

[5] http://hadoop.apache.org

yottabytes (10^{24} bytes). Joking with this, the website states that the 'Utah Data Center was built with future expansion in mind and the ultimate capacity will definitely be "alottabytes"' (NSA, 2018). It also informs that it 'is powered by the massively parallel Cray XC30 supercomputer which is capable of scaling high performance computing (HPC) workloads of more than 100 petaflops or 100,000 trillion calculations each second' (NSA, 2018). 'The steady rise in available computer power and the development of novel computer platforms will enable us to easily turn the huge volume of incoming data into an asset to be exploited, for the good of the nation' (NSA, 2018). The datacentre's automated platforms allegedly can process, in real-time, the 10 gigabytes of Internet traffic inside the U.S.

The Utility of Bulk Data

A recent study conducted with some of the above cited public tools on Google search arguments has concluded that 'search terms can serve as a measure of propensity and can be used to predict the overall proportion of highly qualified [US] Army accessions' (Jahedi, Wenger & Yeung, 2016). This conclusion, by itself, presents an open space for unstructured bulk data use. 'Google's CEO Eric Schmidt admitted as much in 2010: "We know where you are. We know where you've been. We can more or less know what you're thinking about."' (Schneier, 2015)

Also, as shown, from network traffic raw *bulk data* it is relatively easy to extract *bulk metadata*. For intelligence gathering purposes, 'Metadata generally gives us information that we think of as *events* ("A communicated with B at time t")' (GCHQ, 2011:p.10).

A set of bulk metadata databases regarding the protocols exemplified in this article is in use by the GCHQ. The SALAMANCA database contains 'telephone call record *events* from a wide variety of sources'. The full data contains many attributes, but the relevant ones are the *timestamp* and call length along with identifiers as dialled number and caller ID. When involving mobile telephony, the identifiers include 'the IMSI (which is an ID for a SIM card)' and 'the IMEI (which is an ID for a mobile phone handset)' (GCHQ, 2011:p.69). A database called FIVE ALIVE stores bulk IP-IP communications *events*. Each record in it 'summarises a *flow* between two IP addresses', consisting of 'the start of flow time', 'the source and destination IPs and ports and the protocol' and 'optionally extra information on flow size and direction depending upon the protocol' (GCHQ, 2011:p.70). For the HTTP protocol, there is the HRMap database. When entering a web page, an HTTP GET request is generated, always containing the desired URL and often the URL 'of the previously viewed page'. The hostname of the desired URL is the 'HOST', while the hostname of the previous URL is the 'REFERRER'. When considering 'just the hostnames rather than the full URI then this is considered events data' that 'can be viewed as a directed graph of hostnames and is given the name HRMap at GCHQ' (GCHQ, 2011:p.71). One last example is the SKB (Signature Knowledge Base) database, 'for tracking file transfers made on the internet'. An *event* is generated when 'certain file types' are transferred, and the files are identified by their 'format and a hash of some of its content' (GCHQ, 2011:p.71). This format and hash, indeed, is logically similar to the concept of the shape of a data packet used above.

The value of metadata for intelligence gathering is undisputed. Stewart Baker, former NSA General Counsel, has declared: 'Metadata absolutely tells you everything about somebody's life. If you have enough metadata, you don't really need content.' Commenting on these declarations, General Michael

Hayden, former director of the NSA and the CIA, declared 'It is right! We kill people based on metadata.' (Cole, 2014)

HM Government (2016) informs that the analysis of bulk data has: (i) 'played a significant part' in the investigation of 'seven terrorist attack plots disrupted since November 2014'; (ii) 'enabled over 90% of the UK's targeted military operations during the campaign in the south of Afghanistan'; (iii) 'been essential to identifying 95% of the cyber-attacks' over the last 6 months; (iv) been essential in the 'disruption of over 50 paedophiles in the UK in the last three years'.

It is not difficult to conclude that, having these databases of bulk metadata, with the application of *discriminants* (seeds) for filtering (targeting) this metadata, it is possible to collect useful intelligence information.

Nevertheless, *bulk metadata* can also be useful without *discriminants*. The analysis of *bulk metadata* by itself can reveal information not perceivable in the targeted analysis. It can be used to generate large-scale communication graphs showing the relation between phone and Internet users. A tiny number of payphones, say 2 to 5, used exclusively to talk among themselves, could indicate suspicious activity, as the use of communication between members of a terrorist cell. This is a kind of information that emerges from bulk data, with no *discriminants* applied to target data collection. The analysis of bulk data can also reveal unexpected modus operandi ('anomalies and outliers') of the communications pattern. 'Temporal analysis and behavioural pattern-matching can be used to detect hostile network activity from CNE [Computer Network Exploitation] and botnets' (GCHQ, 2011:p.15). Someone using an anonymity protocol like Tor inside a specific network would raise basis for suspicion and further investigation, for instance.

Is Justice the 'last stand' against bulk collection?

Western political science considers the Rule of Law as the shelter that protects citizens' liberties, sometimes against their governments. Hobbes (1651) identified that one would gladly give part of his *liberty* to a government in exchange for some *security*. Locke (1689) identified the paradox posed by the same poles: liberty is worthless without security, but too much security may compromise liberty. Disruptive events can affect a society's disposition regarding this trade-off. An example could be the USA PATRIOT (Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism) Act of 2001, signed into law just 45 days after the 9/11 terrorist attacks. As the name itself explains, many intelligence tools were made available for surveillance, restricting civil liberties.

Nonetheless, the Snowden revelations showed that the U.S. intelligence community was going even further, and implementing surveillance practices and tools not authorised by law (Greenwald, 2014). The analysis of dozens of legal documents issued in the U.S. from 1967 to 2016 showed that the process of increasing surveillance at the expense of liberty started already in the mid-70's (Sivan-sevilla, 2017). The conclusion was that '*the judiciary*, which was able to set the policy tone of restricting government surveillance in the 1970s, has been mostly unsuccessful in influencing the agenda in the same direction since' (Sivan-sevilla, 2017:p.74). Hayden (2016:p.29) explains that the NSA program named Thin Thread was designed for broad collection and processing of metadata from emails, and then adapted to phone calls, and that 'sometime before 9/11' NSA lawyers had argued that 'no system could *legally* do with US data what Thin Thread was designed to do'. Nonetheless, after

Snowden many adjustments have been made to NSA systems and procedures, as also to the law, to at least minimise the gap between theory and practice (Swire, 2015).

The U.S. Congress just approved a new version of the Foreign Intelligence and Surveillance Act (FISA) Section 702, allowing the intelligence community to investigate communications with 'one leg' in the U.S. without warrants for the next six years.

In the U.K. the struggle between the executive and the judiciary persists. In 2015, in a process initiated by two Members of Parliament of opposite parties (one a former Defence Ministry), a High Court stated that the 2014 legislation regarding bulk data retention and surveillance was unlawful, having insufficient privacy safeguards and being contrary to the European Union law (Bowcott, 2015). In 2016, the Investigatory Powers Tribunal, 'the only court that hears complaints against MI5, MI6 and GCHQ', determined that 'the security services operated an illegal regime to collect vast amounts of communications data, tracking individual phone and web use and other confidential personal information, without adequate safeguards or supervision for 17 years', failing to comply with article 8 of the European Convention of Human Rights (ECHR), which protects the right to privacy (Travis, 2016, 2018). The decision considered the 2014 legal framework and did not pose any consideration regarding the Investigatory Powers Act of 2016, which expanded government's bulk powers and tried to retroact ruling lawful previous actions. This was considered an executive 'victory' by some. Then, in January of 2018, the Court of Appeal decided that even the new law is acceptable; investigation must be initiated based on suspicion and authorised by justice.

Simultaneously, and in the same direction, in December 2017 a German court ruled that 'Germany's foreign intelligence agency (BND) must not store the metadata - such as phone numbers - of international phone calls for the purpose of intelligence analysis' (Reuters, 2017).

Conclusion

Bulk collection is relevant to the matters of intelligence for national security and defence. Having the capabilities, both technical and legal, for dealing with massive data can be significant for the foreseeing future. It is comprehensible why intelligence agencies are pushing to have them.

The major technical issue relates to the fact that Internet traffic is still increasing at large rates, due to the inclusion of new users, to the Internet of Things, and to Big Data. To face this problem, however, intelligence services can count on increasing processing power, storage capacity and evolving parallel algorithms and clustering techniques, offering new possibilities for massive data analysis at decreasing costs.

Whether *bulk collection* configures (or not) 'mass surveillance' is subject to debate. Nevertheless, it relates more to privacy in terms of 'freedom of association' than to what someone says in private.

Hobbes, Locke and Rousseau, each with a different perspective, agreed that a portion of individual liberty has to be given in exchange for security. In what measure is a matter to be discussed internally in each society. In this context, it is useful to consider:

> The condition upon which God hath given liberty to man is eternal vigilance; which condition if he break, servitude is at once the consequence of his crime and the punishment of his guilt. (Curran, 1811, "Speech On the Right of Election" on July 10, 1790)

Or, in its shorter version, “The price of freedom is eternal vigilance”. Yet, vigilance is not necessarily a synonym of mass surveillance.

References

Bowcott, O. (2015) 'High court rules data retention and surveillance legislation unlawful', *The Guardian*, 17 July. Available at: http://www.theguardian.com/world/2015/jul/17/data-retention-and-surveillance-legislation-ruled-unlawful.

Cole, D. (2014) '"We kill people based on Metadata"', *NYBooks*, 10 May. Available at: http://www.nybooks.com/daily/2014/05/10/we-kill-people-based-metadata/.

Curran, J. (1811) *Speeches of John Philpot Curran*. New York: Riley.

Dean, J. and Ghemawat, S. (2004) 'MapReduce: Simplified Data Processing on Large Clusters', in *OSDI04: Sixth Symposium on Operating System Design and Implementation*. San Francisco, pp. 137–147.

GCHQ (2011) *HIMR Data Mining Research Problem Book*. GCHQ.

GCHQ (2016) 'GCHQ History'. Available at: http://www.gchq.gov.uk/history/Pages/index.aspx.

Greenwald, G. (2014) *No Place to Hide*. Penguin Books.

Hayden, M. V (2016) *Playing to the Edge*. New York: Penguin Press.

HM Government (2016) *Operational Case for Bulk Powers*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/504187/Operational_Case_for_Bulk_Powers.pdf.

Hobbes, T. (1651) *Leviathan*.

ISO (1998) *ISO/IEC 7498-1 Open Systems Interconnection - Basic Reference Model*. Available at: http://www.ecma-international.org/activities/Communications/TG11/s020269e.pdf.

Jahedi, S., Wenger, J. and Yeung, D. (2016) *Searching for Information Online: Using Big Data to Identify the Concerns of Potential Army Recruits*. Santa Monica. Available at: http://www.rand.org/pubs/research_reports/RR1197.html.

Locke, J. (1689) *Two Treatises of Government*.

Moore, G. and Brock, D. C. (2006) 'Moore's Law at 40 (Chapter 7)', in *Understanding Moore's law: Four decades of innovation*.

NRC (2015) *Bulk collection of signals intelligence: Technical options*. Edited by NRC. Washington: National Academies Press.

NSA (2018) 'NSA Utah Data Center - Serving Our Nation's Intelligence Community', *Nsa.gov1.info*. Available at: https://nsa.gov1.info/utah-data-center/index.html.

Omand, D. (2015) *Understanding digital intelligence and the norms that might govern it*. Available at: https://www.cigionline.org/sites/default/files/gcig_paper_no8.pdf.

Reuters (2017) 'German court rules against foreign intelligence mass communication surveillance loom', *Reuters*, 14 December.

Rid, T. (2015) 'The hype over metadata is a dangerous myth', *Financial Times*, 17 December. Available at: http://www.ft.com/intl/cms/s/0/e0da1c64-a4b5-11e5-a91e-162b86790c58.html#axzz428rXD1BN.

Schneier, B. (2015) 'NSA Doesn't need to spy on your calls to learn your secrets', *Wired*, March. Available at: http://www.wired.com/2015/03/data-and-goliath-nsa-metadata-spying-your-secrets/.

Singh, S. (2000) 'The Code Book: The Secret History of Codes and Code Breaking'. London: Fourth Estate.

Sivan-sevilla, I. (2017) 'Trading Privacy for Security in Cyberspace : A Study Across the Dynamics of US Federal Laws and Regulations Between 1967 and 2016', pp. 73–92.

Swire, P. (2015) *US Surveillance Law, Safe Harbor, and Reforms Since 2013*. Available at: https://fpf.org/wp-content/uploads/2015/12/White-Paper-Swire-US-EU-Surveillance.pdf.

The Economist (2016) 'After Moore's Law', March, pp. 1–37.

Travis, A. (2016) 'UK security agencies unlawfully collected data for 17 years, court rules', *The Guardian*, 17 October. Available at: https://www.theguardian.com/world/2016/oct/17/uk-security-agencies-unlawfully-collected-data-for-decade.

Travis, A. (2018) 'UK mass digital surveillance regime ruled unlawful', *The Guardian*, 30 January, pp. 2–4.

U.S. White House (2014) 'Presidential Policy Directive/PPD-28', pp. 1–9. Available at: https://www.whitehouse.gov/sites/default/files/docs/2014sigint_mem_ppd_rel.pdf.