Privacy with Big Data: A Framework

Completed Research

David Hoffman
Carleton University
DavidHoffman3@cmail.carleton.ca

Abstract

Big data compromises the privacy of individuals and groups. There seems to be no defense but privacy is crucial for personal autonomy and the preservation of democracy. This paper develops the big data and privacy landscape through a literature review and the development of a high level framework. The framework is intended to assist in understanding how the facets of privacy and big data interact at a macro level.

Keywords

Big data, privacy, framework.

Introduction

Data is collected about individuals through: credit card purchases, donations, Internet browsing and searching, libraries, toll booths, travel tickets, tax databases, transportation, employment, insurance, and health care (Katzan, 2010). Daily, society generates 2.5 quintillion bytes of data. By the year 2020 there will be 40 zettabytes of data (IBM, 2018). Statista (2018a) predicts that the Internet will produce 15 exabytes of traffic per year by 2020. The trend is to capture data about everything we do; everyday activities have become “streams of data” to companies (Baruh and Popescu, 2017).

The Internet of Things will lead to a massive number of devices being connected to the Internet by 2020. Statista (2018b) provides a conservative estimate of growth in the number of Internet of Things devices arriving at 30 B devices by the year 2020. Each one of these devices will be generating data and some of it will be personal information. All of this data is being collected, stored, analyzed and used to create knowledge. Just looking at personal data, the knowledge is about us as consumers, patients, clients, customers and products. This is the beginning of the conflict between big data and privacy.

Strang and Sun (2017) conducted a longitudinal literature survey on big data over a span of one hundred years from 1916 to 2016. They identified 79,012 articles from 34 searchable literature indexes which they refined to an analysis set of 13,029 articles. They were able to show that the “big data paradigm” started in 2011 and began to rapidly increase in 2012. Between 2011 and 2016, the research production in the area of big data area included only 2% of the articles on the topic of privacy and security.

Objective

The objective of this paper is to study the extant literature on the interaction of privacy and big data. Then with this understanding develop a general conceptual framework that lays out this interaction in a logical and simple manner in order to understand the landscape of big data and privacy.

The content of the paper includes a discussion of the definitions of big data and privacy followed by an explanation of the methodology. The results of the analysis are discussed and the framework is presented. The paper concludes with limitations and the areas where research and practice can use the framework.

Definition of Big Data

There are different types of definitions used to explain big data. The first comes from the use of the word big. Big data must somehow consume a lot of storage space. Intel stated in 2013 that big data requires a
median data generation volume of 300 terabytes of data per week (Ward, 2013). A second definition focuses on the qualities of the data. Borne (2014) builds on work put forward by Laney (2001) to itemize ten V’s that define the qualities of big data: volume, variety, velocity, veracity, validity, value, variability, venue, vocabulary and vagueness. Volume, Variety and Velocity all seem to be very true today about the ever expanding data available for analysis. The remaining topics are true for all data that might be useful for analysis and is therefore not expressly applicable to big data. Yet another way to define big data is by the tools required to analyze it. The assumption is that big data requires the development of new heavy weight analytical and storage technologies (Ward, 2013). Big data in its new form requires a paradigm shift in storage and analytical tools in order to handle the new forms of data in a timely, usable manner (NIST, 2015).

Conceptually big data is unique; it is the synthesis of all of the above definitions. Big data is defined as any data set that requires specialized tools and analytic methods to reap new insights that can only be generated with significantly large and diverse content.

**Definition of Privacy**

A person’s property includes “every form of possession – intangible, as well as tangible” (Warren and Brandeis, 1890). They argue that common law provides the right to the individual to decide for themselves what aspects of their life will be shared and with whom. Floridi (2005) agrees, when people talk about “my information” they are not talking about something like “my car” but it is more “my body”. His point is that information about us is personal and is part of what defines us as people. It cannot be separated and given away or sold then it is gone. The marketplace has managed to find a way to commoditize privacy by trading personalized services for information about the individual using the service (Walsh, Parisi, and Passerini, 2017). People give away or sell their privacy possibly without even realizing what they have done.

Humans are unable to develop their personality properly without a secure private place (Boehme-Neßler, 2016). Part of this is to decide what information about themselves they are willing to share. Privacy is not only a legal construct; it is also a “psychological and anthropological necessity.” People need to control access to themselves. Without this autonomy democracy could not function.

Privacy is not security. Gahi, Guennoun and Mouftah (2016) look at the technological handling of big data and describe privacy as the need to handle sensitive data with the care that it requires. They state that data related to privacy of individuals must be protected differently than other data. Security is the protection of the data from individuals who do not have a right or need to access the data. Privacy focuses on “the use and governance” of personal information (Jain, Gyanchandani and Khare, 2016).

Privacy can then be defined as the right of people to control access to their lives including access to information about themselves. This data cannot be separated from themselves because it is an integral part of their being. The right to privacy includes the right to be left alone. External agencies that have access to personal information have an obligation to protect and preserve the information much like they would protect the person to whom it belongs. There is also a need to be proactive when dealing with new technologies in order to ensure the fundamental concepts of privacy are not violated.

**Method**

With the objective of developing a framework on privacy and big data interaction, Schwarz, Mehta, Johnson and Chin (2007) was used as a guide to ensure rigor in the method.

The search for articles to include in the study was performed in four steps and is summarized in Table 1. The goal was to examine a more focused set of literature over a larger period of time from a management and information systems perspective and to also include the current thinking from all sources. The choice of date ranges was made with this goal in mind. Business and Information System journals (the top six on the ABDC list) were searched over a ten year period. All journals over the previous year were searched. All online searches were made against the whole article – title, abstract, key words and the body.

The goal of the review process was to identify articles that had a management impact or could provide insight to managers and academia on the subject of big data and privacy as they interact with each other.
The first step in the review process was to read the titles, abstracts and key words to determine whether the article was related to the topic of big data and privacy. Articles would be excluded at this point if they did not use these key words or were deemed to not be directly relevant to the objectives. This step would specifically exclude articles that related to the implementation of privacy preserving technologies and articles that related to the perceptions of individuals such as trust.

<table>
<thead>
<tr>
<th>Step</th>
<th>Source</th>
<th>Search Criteria</th>
<th>Start Date</th>
<th>End Date</th>
<th>No. of Articles</th>
<th>Excluded</th>
<th>Total Reviewed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Any journal</td>
<td>Big, Data and Privacy</td>
<td>1 Jan 2016</td>
<td>15 May 2017</td>
<td>30</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>Any journal</td>
<td>Data and Privacy</td>
<td>1 Jan 2016</td>
<td>15 May 2017</td>
<td>15</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Business journals</td>
<td>“Big Data” and Privacy</td>
<td>1 Jan 2007</td>
<td>15 May 2017</td>
<td>37</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Totals</td>
<td></td>
<td></td>
<td></td>
<td>110</td>
<td>61</td>
<td>49</td>
</tr>
</tbody>
</table>

Each article was then read in detail looking for common threads relating to privacy and big data that are relevant to the objective of the analysis. The goal was to find the thought or perspective that dominated the thinking in the article on the subjects of privacy or big data. These ideas were first written out using the terms of each article and later consolidated into a smaller subset of themes. A facet is a definable aspect that makes up a subject (Merriam-Webster). These common themes or threads will be referred to as facets henceforth. Table 2 summarizes the intersection of the facets of privacy and big data addressed in each of the included articles.
Table 2 – Cross Reference by Articles of Privacy and Big Data

<table>
<thead>
<tr>
<th>Facets of Big Data</th>
<th>Active / Passive</th>
<th>Commercialization of Data</th>
<th>Data Aggregation</th>
<th>Datafication</th>
<th>Knowledge Creation</th>
<th>Re-Identification</th>
<th>Security</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consent</td>
<td>7</td>
<td>5</td>
<td>13</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>52</td>
</tr>
<tr>
<td>Privacy By Design</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Privacy Preserving</td>
<td></td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>36</td>
</tr>
<tr>
<td>Rights and Ethics</td>
<td></td>
<td>7</td>
<td>9</td>
<td>17</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>17</td>
<td>37</td>
<td>17</td>
<td>24</td>
<td>25</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>

Analysis

Consent

Consent is required to collect and use the data and must be maintained for all uses of the data (Nicholson, 2016; Bietz, Bloss, Calvert, Godino, Gregory, Claffey, Sheehan and Patrick, 2016) however valid consent is not open ended and does not extend to third parties (Pascalev, 2017). Pascalev (2017) states, the failure of notice and consent is widely acknowledged. Part of the problem lies in the fact that notice and choice puts the onus on the individual to know and understand their risks (Baruh and Popescu, 2017). Putting the burden on the individual also requires that the lay person must be able to read the notice and consent form in their available time and at their language ability (Pascalev, 2017).

Accompanying the problems associated with individuals being able to make good decisions about consent regarding their personal information is corporate interests. One misconception is that consumers think the user agreements they agree to online constitute informed consent or the notices are not read. Social media privacy controls are not informed consent; they are user agreements to access the site and reversible at any time by the provider (Frizzo-Barker et al, 2016; Pascalev, 2017; Dai, Qiu, Qiu, Chen and Wu, 2017). People agree to user agreements and provide personal information because they want to receive customized, personal services and the providers will refuse service without access to personal information (van de Pas and van Bussel, 2015). People also don’t understand what should be private and as a result don’t value their personal information in order to have an awareness of how to handle it properly. This is a key milestone in addressing privacy issues (Frizzo-Barker et al, 2016; Sidgman and Crompton, 2016; Buck, Horbel, Kessler and Germelmann, 2014).

As the use of big data and analytics becomes more common people will likely have been found to have not consented to the use of personal information for some analytics (Nicholson, 2016). Problems with informed consent are caused by unforeseen uses of the data (Frizzo-Barker et al, 2016; Ohlhausen, 2014; Moorthy, Labiri, Biswas, Sanyal, Ranjan, Nanath, and Ghosh, 2015; Kum, Krishnamurthy, Machanavajjhala, Reiter and Ahalt, 2014). This includes the creation of new, undisclosed personal information or insights using analytics that conflict with consent and collection notices agreed to by users (Nicholson, 2016; Sax, 2016). Also, “the integrity of the subject’s identity is breached when data or information is added to subject’s identity without consent” (Mittelstadt, 2017). The net results is that as our lives become datafied there becomes very little that is hidden from the consumers of our personal information (Martin, 2016).
Another way that informed consent becomes irrelevant is in the use of big data and analytics to identify and describe groups. Barocas and Nissenbaum (2014) identify the “tyranny of the minority”, a situation where a few people who share information about themselves, may disclose information about a group of people who share the same observable traits but are unwilling to share their personal information. These groups have privacy interests just like individuals because the patterns used to group people then function as identifiers threatening individual privacy. All members of an ad hoc group are expected to respond similarly (Mittelstadt, 2017). Thus ability to infer “rules” from a select group of people that generalize to an entire population renders consent irrelevant. The goal is a representative sample.

Consent is an activity that is performed before any data is collected. Its purpose is to inform the data owner about what data will be collected and how it will be used. Consent is used to inform downstream users of the data about how they can and cannot use the data.

**Privacy By Design**

Privacy by Design is based on the premise that privacy awareness is a system development issue. The first step is a ‘privacy first’ culture. This involves “proactive privacy protection in privacy conscious, data rich environments” (Everson, 2016). Privacy by architecture aims to reduce privacy risk in system development early in the design cycle (van de Pas and van Bussel, 2015).

The first goal of privacy by design is to safeguard the data from unwanted disclosure, or apply security measures, then extract knowledge without compromising privacy (Jain et al, 2016). Data security does not lead to privacy regulation compliance or to privacy aware applications (van de Pas and van Bussel, 2015). New health care systems need to be designed with security and privacy built in using a patient centric model. The patient defines the security policies for their own data and all uses including big data analytics (Blobel, Lopez and Gonzalez, 2016).

Storing data electronically brings risk from hacking or unauthorized access (Ohlhausen, 2014; Nunan and Di Domenico, 2013). If there is a data breach it doesn’t matter who holds your data and how well it is protected (Ohlhausen, 2014) however data encryption requires too much overhead for big data applications (Fang, Wen, Zheng and Zhou, 2016).

The architecture of mobile devices enables the linking of data to create deep views into people’s lives. Smart phone applications quite often access the following data types whether required or not: location, address book, photos, or the unique phone ID. The unique phone ID or IMEI / UDID acts as an identifier for all other data harvested from a specific phone by all applications (Dai et al, 2017). It is unlikely that phone owners are aware that the phone ID is being taken and used to link their data across applications. All of the data from various applications is easily linked to an identifiable person using the unique phone ID making data identification easier across data sets.

Privacy by Design must restrict the data being captured to only that which is necessary for the current application and that data which identifies the subject or is sensitive must be properly protected to ensure the individuals privacy. Privacy by Design must also give consideration to how data may be linked with other data in order to avoid disclosing personal information through the aggregation into big data.

**Privacy Preserving**

Privacy preserving is a global term for how organizations handle data about people in order to ensure the individuals privacy. There are three main streams of privacy preserving: anonymization, de-identification, and encryption (Fang, Wen, Zheng and Zhou, 2016). Data encryption is mainly focused on protecting the data from users that should not be seeing it so it is considered part of security. The focus of privacy preserving then is to protect the privacy of individuals given that the data is already available in a usable format.

Fang, Wen, Zheng and Zhou (2016) conducted a survey of various privacy preserving technologies and methods. Their analysis recognizes that the current privacy preserving methods are still immature. They also identified that aggregating various data sets into big data can circumvent privacy preserving techniques. They conclude their analysis with the thought that in order to solve privacy preserving problems there is a need for legal constraints.
Anonymization and de-identification represent two different end uses for the data. Anonymization is the process of rendering data sets anonymous by grouping records, generalizing the data, suppressing specific data or adding fake data records to the set in order to maintain the analytic usefulness of the data set (Jain et al, 2016). De-identification is the decoupling of identifiers and quasi-identifiers from sensitive data, while maintaining each individual record (Daries, Reich, Waldo, Young, Whittinghill, Ho, Seaton and Chuang, 2014).

Data anonymization is a compromise between the value of the data for processing and the amount of sensitive data released (Fang, Wen, Zheng and Zhou, 2016; Daries et al, 2014; Jain et al 2016). Kim, Lee and Chung (2017) studied the use of privacy-preserving data cubes built with anonymized data. They discovered that as cube generalization increases there is an information loss penalty. As data streams are getting larger, anonymizing them is becoming more challenging. Anonymization does not appear to be a viable way to ensure data privacy (Jain et al, 2016).

Sufficient de-identification removes statutory elements and all quasi-identifiers. Quasi-identifiers enable identification in combination with other data or other data sources (Daries et al, 2014). De-identification is difficult to achieve and unable to defend against the various different threats from big data (Barocas and Nissenbaum, 2014; Angiuli, Blitzstein and Waldo, 2015).

The availability of big data increases the risk of re-identification using homogeneous sensitive data. Volume and variety also increase the risk of re-identification. Privacy preserving anonymization is insufficient for big data analytics and encryption is inefficient (Jain et al, 2016). Despite the efforts to secure privacy with anonymization, big data enables populations to be analyzed for patterns and then people are put into populations based on observable facts (Barocas and Nissenbaum, 2014). “It is often surprising how little information is needed to re-identify a subject” (Angiuli et al, 2015).

The concern is not the collection of data but rather its aggregation, correlation and de-anonymization (Pascalev, 2017; Nicholson, 2016; Spiekerman et al, 2015). Combining data sets that previously had no privacy implications can suddenly create a privacy problem (Nunan and Di Domenico, 2013; Nicholson, 2016). Data science and information technology advances are destroying old assumptions about anonymity (Kulynych and Greely, 2016).

**Rights and Ethics**

The current ethical question is not whether to collect data because it is being collected but when and how to analyze the data, what to look for, and what questions to ask (Mai, 2016). Data owners have a legal right to privacy, including information privacy, and the expectation that their data will be used appropriately (Dinev, 2014; Spiekerman, Acquisti, Bohme and Hui, 2015). There are three ethical practices that need to be addressed: the right to be forgotten, the right to data expiry, and the ownership of social graphs (Nunan and Di Domenico, 2013).

One challenge to the ethical standards listed above is the threat of government use of surveillance and public technologies combined with predictive analytics to monitor citizens (Power, 2016). Preventing a totalitarian state will require vigilance; democracy needs individual rights, civil liberties, and privacy combined with self-government, safety and economic opportunity (Boehme-Neßler 2016; Power, 2016). One problem is how much privacy are people willing to give up for protection from terrorists and crime (Ring, 2016; Katzan, 2010).

People are treated as data subjects with digital identities (Spiekerman et al, 2015; Walsh et al, 2017). People treat their privacy as a commodity to be traded for access to desired services. Unfortunately, treating privacy as a commodity makes it a disposable good rather than a universal right (Walsh et al, 2017). The trend is to capture data about everything we do, everyday activities are or have become “streams of data” to companies (Barocas and Nissenbaum, 2014;Baruh and Popescu, 2017). Data capture has become so easy that some people believe we should capture everything to reduce crime and terrorism (Power, 2016). In fact, government surveillance has become normal and yet Internet companies might be a bigger threat than government (Ring, 2016).

Companies collect more data than they should and create more knowledge than users imagine creating privacy and security violations (Gahi et al, 2016). Analysts are continuously working to create knowledge that is not immediately evident from the total set of aggregated data. The business model of Internet
companies is based on the notion that they may use whatever insights they generate out of the data available. “Big data’s entrepreneurial potential resides in the fact that advanced data mining techniques can extract / generate unanticipated, non-trivial, new, and (commercially) interesting insights” (Sax, 2016; Barocas and Nissenbaum, 2014).

**Framework**

Rights and Ethics and Privacy are not data constructs but fundamental needs of society. They exist regardless of big data. Rights are dependent on our legal entitlement to privacy and the security of people including personal information. Ethics guides the use of personal information in order to ensure that people are treated justly and appropriately. Privacy is an understanding of individual and corporate needs to be able to operate autonomously without external influences caused by a knowledge of personal information.

Consent and Privacy by Design are activities that must occur before any data is collected. Consent is the agreement between the data owner and the data gatherer regarding what data is collected and how it will be used. Privacy by Design is the effort the data gatherer takes up front to ensure all personal data is properly protected and utilized.

![Privacy Big Data Framework](Image)

Security is a perimeter that protects all aspects of the gathering, storing, using and disposing of data. It keeps out the external parties that would steal, change or destroy data. It also protects the data from
unauthorized access from inside where analysts may want to exploit the data contrary to the direction of the organization holding the data. It is a protective layer that allows the data to be used appropriately and with the privacy interests of the subjects intact.

Active data is entered directly by the data owner while passive data is collected with or without the user’s awareness but without direct action by the user. Datafication is the rampant collection of any all data available whether there is an immediate use for the data or not.

Once the data is collected the data must be protected in order to preserve the privacy of the data owner. The two main methods of preserving privacy are anonymization and de-identification. There are problems with both methods when big data is being used for analysis.

Data uses after data collection reveals the purposes and benefits of big data. Knowledge creation can be a noble or an ignoble activity. Big data can find cures for cancer, identify terrorists or provide personalized services. It can also identify behaviour patterns that can be exploited for corporate gain. Data analytics and mining can also be used to create information about people that previously did not exist but is highly private. Data aggregation is the relentless pursuit of more data sets about people that can then be exploited for gain. Data brokers are the usual custodians of this data but government also amasses great quantities of data about people. The commercialization of data is related to data aggregation as disparate data sets are analyzed to understand behaviour patterns that can then be monetized. Raw data may also be sold to other entities. Re-identification is the process of using big data to analyze data sets that have been treated for privacy protection with the sole goal of obtaining access to identified sensitive data.

**Contribution**

The framework is useful to inform academia and practice of the various facets of big data and privacy. It is designed to provide a way to look at the environment and understand the process of big data creation with consideration of privacy. It also provides insight into the fact that privacy considerations are not a one time, up front decision at the point of data collection.

**Future Research**

One major stand out in the development of the framework is that data owners don’t fully understand what they are agreeing to when they sign consent forms or end user agreements for services. This highlights a need for research and development of tools to help end users make better informed decisions about their personal data. This research could also be used to expose bad actors who gather data inappropriately.

There is also a need to develop security, encryption and privacy preserving methods that meet the performance requirements of big data. Finally, research into new methods of data anonymization and de-identification that can resist de-identification by big data while preserving the usefulness of the data for analysis are required.

**Implications for Practice**

The framework highlights the need for a long term view of data privacy when collecting data. Corporations need to adopt internal policies and procedures that will ensure correct handling of sensitive data. The legal environment is changing and corporations must be aware of the requirements in every locale where they do business and where their clients reside. The framework provides a holistic view of big data and privacy that will help practice organize their thinking to handle personal information appropriately.

**Limitations**

The framework is designed to understand the big picture of big data and privacy. As such it is not useful for interpreting the capabilities of specific tools to ensure privacy. Likewise, there are no methods or metrics for determining levels of privacy or strategies to protect privacy. While the framework has been developed in a general manner it should be reviewed periodically to ensure that new technological capabilities and legal requirements have not generated changes that should be incorporated.
Conclusion

At every turn it seems that big data or bigger data is able to defeat every attempt to prevent our privacy from leaking all over the Internet, corporate offices and government surveillance agencies. It is time for all parties to work together to find a solution to protect individual privacy and the future of democracy. We can argue that it is too late and that may be true but if no one acts then we will eventually reach a point beyond recovery.

Bibliography