Big Data Analytics: Balancing Individuals' Privacy Rights and Business Interests.

by

Columba James Jordan

Canterbury Christ Church University

Thesis submitted
for the degree of MSc by Research

2017

ABSTRACT

This research thesis analyses and discusses the importance of having a legal framework that can control and manage the use of data during the Big Data analysis process.

The thesis firstly examines the data analytics technologies, such as Hadoop Distributed File System (HDFS) and the technologies that are used to protect data during the analytics process. Then there is an examination of the legal principles that are part of the new General Data Protection Regulation (GDPR), and the other laws that are in place in order to manage the new era of Big Data analytics. Both the legal principles Chapter and data analytics Chapter are part of the literature review.

The IT section of the literature review begins with an analysis of the data analytics technologies, such as HDFS and Map-Reduce. The second part consists of the technologies to protect privacy, especially with respect to protection during the data generation phase. Furthermore, there is a discussion on whether these current technologies are good enough to provide protection for personal data in the Big Data age.

The legal section of the literature review starts by discussing some risk mitigation schemes that can be used to help individuals protect their data. This is followed by an analysis of consent issues in the Big Data era and later by an examination of the important legal principles that can help to control the Big Data process and ultimately protect individuals' personal data.

The motivation for carrying out this research was to examine how Big Data could have an effect on ordinary individuals, specifically with respect to how their data and privacy could be infringed during the data analytics process. This was done by bringing together the Big Data worlds from the legal and technological perspective. Also, by hearing the thoughts and views of those individuals who could be affected, and hearing from the experts who could shine a light on the realities in the Big Data era.

The research includes the analysis and results of three surveys, constituting over 100 respondents, who expressed their views on a number of issues, including their fears about privacy online. This included a survey of mainly closed questions for students at Canterbury Christ Church University, a survey monkey survey for students at University College Cork, in Ireland and finally a survey for students in Sri Lanka.

Questions were posed to some experts in areas of IT law and Big Data analytics and security. The results of these interviews were analysed and discussed, producing much debate with respect to what can be done to manage and protect citizens' personal data privacy in the age of Big Data analytics. The software packages Statistical Package for the Social Sciences (SPSS) and Minitab were used to analyse the results of the surveys, while Qualitative Data Analysis Miner (QDA miner) software was used to analyse the results of the interviews.

ACKNOWLEDGEMENTS

I want to thank my Academic Supervisor Dr. Man Qi and my Supervisor chairman, Dr. Abhaya Induruwa, Canterbury Christ Church University, for their help during this research project. I also want to thank Dr. Gowri Nanayakkara, Canterbury Christ Church University, for her help with the legal side of the thesis and other advice that she offered me.

Thanks also to Dr. Chris Harvey and Dr. Sabina Hulbert, Canterbury Christ Church University, for their help and advice regarding the use of the Bristol Online Survey (BOS) software at the University. Dr. Harvey was also very helpful when he explained and showed me how to use the Minitab analytical software, which I used to analyse the survey results.

I also have to thank Ms. Alison Knight, Research Fellow at Southampton Law School, University of Southampton, who has been very helpful to me during the course of my

research, she also very kindly put me in contact with other academics in the research area of Big Data law.

Thanks also to the following for their help with my research;

Dr. Henry Pearce	Lecturer in Law at University of Hertfordshire
Professor Eerke Boiten	Professor in Cyber Security at De Montfort University
Professor Van Der Sloot	Tilburg Institute for Law, Technology, and Society
Dr. Anne Alexander	Co-ordinator, Cambridge Digital Humanities Network

Acronyms

Art	Article
BOS	Bristol Online Survey
BUG	Bottom Up Generalisation
DBMS	Database Management Systems
DPA	Data Protection Act
ECHR	European Convention on Human Rights
E – Commerce	Electronic Commerce
EU	European Union
GDPR	General Data Protection Regulation
HADOOP	Highly Archived Distributed Object Oriented Programming
HDFS	Hadoop Distributed File System
HQL	Hive Query Language
ICO	Information Commissioner's Office
No SQL	Non Relational Structured Query Language
PPDM	Privacy Preserving Data Mining
QDA Miner	Qualitative Data Analysis Miner
SM	Survey Monkey
SPSS	Statistical Package for the Social Sciences.
SQL	Structured Query Language
TDS	Top Down Specialisation
TSHC	Trusted Scheme for Hadoop Cluster
UK	United Kingdom

WP Working Party

LIST OF FIGURES

TABLE OF CONTENTS

1. CHAPTER ONE-INTRODUCTION

1.1. MAIN INTRODUCTION

The Big Data world is upon us, where large volumes of data are collected and processed in real time, while the data is used by organisations, who wish to be innovative and forward thinking. The source of this data ranges from Internet of things sensing devices to social media applications. The advancement of computer processing power has resulted in these companies reusing data in order to extract the many benefits it offers. These include assisting in the prediction of climate change and the likelihood of an epidemic spreading. However, the Big Data analytics process has raised concerns regarding the negative effects it has on individuals' privacy and the protection of their personal data.

The typical definition of Big Data is the three Vs: volume (consisting of large amounts of data), velocity (created in real-time) and variety (being structured, semi-structured and unstructured). As previously mentioned, the perceived strength of Big Data is that the analysis of Big Data will enable more accurate identification of a consumer's characteristics than the more traditional marketing methods. (Kitchin et al, 2016) The vast majority of literature and experts in Big Data, form the same view that the definition of Big Data is the three Vs. For example, Harry Pence of the State University of New York, Rob Kitchin, Ralph Schroeder of the Oxford Internet Institute and Donna Burbank of the Global Data Strategy form this view.

As this technology advances at an ever increasing pace, it is necessary to develop a happy medium, where the innovation will not be prevented from occurring, while simultaneously providing individuals with the appropriate privacy protection measures.

The thesis examines the current technologies that analyse data and the respective technologies that can protect privacy, and asks if they are good enough to help protect individuals' data.

The current legal framework is then examined to determine if it is likely to provide the legal safeguards for personal data privacy, which is a challenge as the technology is moving at a faster pace than the law.

The argument has been made for privacy by design to be implemented in Big Data analytics technologies, but this is not as straight forward as some would like to believe. One method of achieving this involves minimising the data. This means that the personal data used should be keep to the smallest amount possible to achieve the objective of the data.

The discussion on privacy technologies involves the use of anonymisation and encryption methods to protect individuals' personal data. The question is posed whether these technologies are good enough to properly protect personal data.

For instance, anonymising static and structured data can be problematic, as there can be issues regarding proper comparability and verifiability. Big Data analytics brings about different problems to these data properties as the data formats increase to unpredictable and unstructured flows.

The thesis gives an overview of the data protections laws that are presently in place, there is a more detailed examination of the new General Data Protection Regulation. It is hoped that this regulation will provide a strong basis for data protection to be achieved in the Big Data age. In particular the regulation sets a high standard for the data controllers to live up to with respect to consent. As the consent provided by the data subject must be deemed to be unambiguous where the data is not sensitive, and explicit where the data is sensitive. These new higher standards for consent are to be welcomed, but how this can work in practise is another issue. Indeed, some experts believe that consent is an obsolete notion, they do suggest that there should be a form of granular consent and one that an individual can comprehend.

The new regulation also ensures through the purpose limitation principle, that data must only be collected if there is a specific and appropriate purpose for its use. This principle is necessary in order to protect personal data, but there is an argument that it may have a negative impact on Big Data analytics.

The data minimisation principle is another component of the GDPR, and it demands that companies reduce the quantity of data that is collected and processed. Furthermore, the amount of data collected should not exceed what is required in order to achieve the aim of the collected data. These issues will be analysed in the thesis.

There is breakdown of the research methodologies used to complete the thesis and others that were not required, this includes both the quantitative methods and the qualitative methods. The quantitative method that was used was inferential statistics, and it was used to make an inference about the data that was generated from the surveys.

The qualitative methods, which were used during the course of the research, included questionnaires and interviews. The questionnaire method was required for the three surveys and the interviews and interview method was required for the interviews with the IT law and privacy technology experts.

As part of the research, a number of surveys were conducted in which students were asked to give their opinions on a number of issues related to users' privacy online. Some of these included asking individuals if their explicit approval should be sought before their data are collected and asking how they feel when their data is collected by public bodies. The surveys provide a valuable insight into what ordinary members of society feel about data privacy online. There is a detailed section consisting of the survey results and analysis, which includes the use of SPSS and the Minitab software packages.

The results and analysis Chapter also includes the outcome of the interviews, which were carried out with experts in IT law and privacy technology. These results help to shine a light on how the Big Data analysis phenomenon can be managed properly, for both the ordinary individual and organisations.

1.2. AIMS OF THE THESIS

Examine the Big Data analytics technologies and the respective Privacy Protection Methods.

Examine how the General Data Protection Regulation is going to protect individuals' personal data, and the importance of the legitimate interest condition, the concept of consent and the data minimisation principle.

Analyse the results of the surveys, using the SPSS and Minitab software applications, in order to hear the opinions of ordinary individuals with respect to their online privacy.

Examine the interview responses from a number of experts, in the fields of IT law and privacy technology, using the QDA Miner software.

1.3. THESIS CONTENTS

The thesis begins with the literature review which is Chapter two. The first section of the literature review, 2.1.2, examines the technologies that are used for data analytics, such as HDFS and Map-Reduce. This is followed by an analysis of the technologies to protect privacy and Big Data privacy in the data generation phase, section 2.1.3.

Finally, in the first section 2.1.4, there is a discussion about whether the technologies to protect privacy are strong enough to fulfil their purpose as the Big Data analytics processes continue to grow exponentially.

The second part of the literature review Chapter, section 2.2.2, analyses the legal principles and framework that exists, in order to protect individuals' privacy, and mitigation schemes, which can help social media users and others to protect their personal data.

The legal section continues to examine the important legal principles that are part of the new GDPR.

This includes a discussion of the purpose limitation principle, which is part of the GDPR, where data can be collected if there is a specific and legitimate purpose. Then the data minimisation principle is analysed, this principle stipulates that companies must reduce the quantity of data that is collected and processed, and indeed not collect any more than is required in order to realise the exact aim of the collected data. Finally, the concept of consent is examined, this involves viewing the relationship between consent and the new GDPR and the complexities involved

Chapter 3, the research methods Chapter, will identify the appropriate quantitative and qualitative research methods that were used during the course of the research process.

Chapter 4 will describe and highlight the results and analysis of the surveys and the interviews.

2. CHAPTER TWO - LITERATURE REVIEW

2.1. LITERATURE REVIEW ON THE TECHNOLOGIES FOR BIG DATA ANALYSIS AND PRIVACY PROTECTION

2.1.1. INTRODUCTION

The dramatic rise in the use of modern technological devices, such as the iPhone, iPad and other devices, has ensured that large volumes of data are produced as a result of commercial and social media activities. The levels of data production are further enhanced, due to data being collected from sensors and networks and the digitalisation of the processes involved in the use of these devices.

It is estimated that the quantity of data generated globally is increasing by one hundred per cent every two years. In order to put the data levels into some context, the data is likely to increase from four and a half zettabytes (four and a half trillion gigabytes) in 2013 to forty four zettabytes by 2020. (European Parliamentary Research Service, 2016, P2)

To further gain a realisation of these data levels, it is also estimated that thirty billion portions of content are shared on Facebook every month, there are twenty billion Internet searches every month. Additionally, there are more than seventy two hours of video data uploaded to You-Tube every minute, Twitter users create 277,000 tweets and millions of networked sensors linked to mobile phones, which create this vast amounts of data. (Bhala et al, 2016, P469)

The mass collection of data, which is known as Big Data, can be collected from Internet communications, electronic commerce (e-commerce) activities, e-government, mobile apps, social media and sensors in items connected to the Internet of Things. The Big Data process is gaining further momentum, as technological advancements continue unabated, particularly in respect of the reduction in storage expenditure, greater networking capabilities, advanced analytical software and additionally the accessibility to Cloud computing services. As a result of these technological innovations, large volumes of data can be stored and processed very efficiently and rapidly. (European Parliamentary Research Service, 2016, P2) Data analytics processes enable the analysis of these large datasets, which results in the identification of patterns and relationships. The patterns and relationships can facilitate the process whereby information and real facts are garnered from the unprocessed data. Subsequently, this real and processed data can be employed to assist in the development of new innovations and enhance decision making processes.

This Chapter, which is the IT section of the literature review, examines the important privacy technologies and issues that will have a bearing on Big Data analytics, presently and in the immediate future.

The first part of the Chapter examines the technologies for data analytics, specifically the HDFS and its main features. There is then an analysis of the operation of the HDFS and Map-Reduce. (Lydia et al, 2016, P100)

The Highly Archived Distributed Object Oriented Programming (HADOOP) system involves the use of two essential parts, firstly the HDFS, which deals with data storage. Secondly, Map-Reduce, which manages data processing. There is also a discussion of the Big Data database systems, which are associated with Hadoop.

The second part of the Chapter focuses on the technologies to protect privacy, this includes an examination of infrastructure security, data privacy and data management with respect to Big Data analytics. Next there is an analysis of Big Data privacy in the data generation phase, which includes a discussion of the measures to restrict access and falsify data, in order to protect the data and Big Data privacy at the data storage stage.
Finally, there is an analysis of whether the technologies, such as Hadoop, are good enough to efficiently protect privacy, this includes a discussion on the limitations of the Hadoop system software. Also, there is an examination of the privacy preserving data mining method algorithms and the limitations of anonymisation.

2.1.2. TECHNOLOGIES FOR DATA ANALYTICS

Big Data consists of structured, semi-structured and unstructured data, which are vast and complex. As a result, it can prove to be problematic when processing data using conventional data processing software. Also, the traditional relational Database Management Systems (DBMS) and statistics software do not provide an adequate platform with which to process Big Data. In order for these platforms to be of use, there would have to be numerous corresponding software packages operating on numerous servers, which would not be practical. (Lydia et al, 2016, P100) But, the Big Data analysis can be carried out on the Hadoop software package, which also consists of its associated tools such as Map-Reduce, Apache Hive and Spark and Non Relational Structured Query Language (No SQL) database systems.

HADOOP DISTRIBUTED FILE SYSTEM

Hadoop Highly Archived Distributed Object Oriented Programming, is the main software package used for structuring Big Data, while it also ensures that the data can be used for analytics purposes. It is an open source framework software, which assists in the storage, processing and attainment of valuable information from Big Data.

The main features of Apache Hadoop are as follows;

Firstly it is **Scalable;** Commodity computers can be added and will not alter the data formats, such as the way in which data is compiled.

Secondly**,** it is very **Cost effective**; Hadoop enables many parallel computing operations to occur on the working nodes, and the resulting output is much less costly, thus it is economically possible to process all the data.

It is **Flexible**; Hadoop is able to process any form of data, whether they are structured or unstructured and even if they come from several sources.

Finally Apache Hadoop is **Fault tolerant**; if the connection to one storage device is broken, Hadoop automatically resends the processed data to a different storage device and the Hadoop system continues to process data.

HDFS AND MAP-REDUCE

The Hadoop system involves the use of two essential parts, firstly the Hadoop Distributed File System, which deals with storing the data. Secondly, Map-Reduce manages the processing of the data. But, the Hadoop system also consists of the following elements, which are Hive, Pig latin, Mahout, Apache Oozie, Hbase, Flume and finally Sqoop.

The Hadoop system is a Java software file system, and Hadoop uses hash functions to deal with data elements using keys and values.

HDFS consists of a master node and numerous data nodes, also known as slave nodes. The master node performs the file system tasks such as renaming, opening and closing files, regulating the client's access to files and also controlling the file system namespace. (Tutorialspoint, 2017)

The data nodes are responsible for the data storage operations of the system. These nodes execute, read and write tasks on the file systems, when there is a request from the client. Also, they carry out tasks such as block creation, deletion and replication in accordance with the instructions from the master node.

The HDFS system allows for the data to come in, it then separates the data into specific parts and sends these parts to others storage devices in a group, this then enables parallel processing to occur. To ensure that the system is fault tolerant, HDFS duplicates every part of data three times and then sends the copied data parts to specific nodes or storage devices, simultaneously putting a copy of the data part on a different data node on each occasion. Consequently, data which are on nodes or storage devices that crash can be located in a different place within a group. This enables processing to carry on even when the malfunction is being dealt with.

Every group contains a single Name Node that controls the file system functions and the supporting Data Nodes control the data storage on the specific compute nodes. (searchbusinessanalytics.techtarget.com, 2017)

Hadoop Map-Reduce is part of the overall Hadoop software infrastructure, which consists of applications that process large amounts of data. This occurs as a result of many parallel computing operations that occur on the data nodes.

The Map-Reduce software performs two functions, in essence it maps the data and reduces it. Map separates data and then segments the data into a set of data elements. Reduce tasks gets the output data from a number of map tasks. The data input and output are both contained in a secure file system. The reduce functionality is executed after the map action. (Tutorialspoint, 2017)

To provide further clarity, Map-Reduce makes use of the Hadoop Distributed File System. Using a Hive Query Language query (HQL query), this is sent by the user, the master storage device uses the map system to designate sections of this specific query to the data storage device to execute. The data device then executes these designated sections and sends back the specific results to the master device. The master device subsequently

decreases these specific results into an organised result, which is sent back to the user who carried out the HQL query. (GU, 2014)

Map-Reduce utilises HDFS. This occurs when an individual submits a HQL query, the master node then employs a mapping procedure to designate elements of the query to data nodes in order to be implemented. The data nodes implement their designated section and deliver their specific results to the master node. The master node aggregates these specific results into an interrelated result, which is delivered to the individual who carried out the HQL query.

BIG DATA DATABASE SYSTEMS

HIVE

Apache Hive is a data warehouse software system, which is used to summarise, query and analyse the datasets in Hadoop files.

Hive has a Structured Query Language (SQL) type of interface with queries written in the Hive Query Language, which can be used to query the data that is located in different databases and file systems that are part of the Hadoop framework.

Furthermore, the Hive Query Language can execute Map-Reduce scripts, which can be attached to queries. (searchdatamanagement.techtarget.com, 2017)

NO SQL OR NON RELATIONAL SQL

No SQL is a group of Database Management Systems that do not adhere to the full set of parameters of the relational Database Management Systems, while querying data is not possible by using the regular SQL language. No-SQL type system software is normally applied in the operation of Big Databases, and especially those which are likely to suffer from operational issues, due to the limitations of SQL and the relational databases. (techopedia.com, 2017)

Figure 1 shows the architecture of a system, in which both the windows application and the web application are able to perform computation on the Hadoop framework.

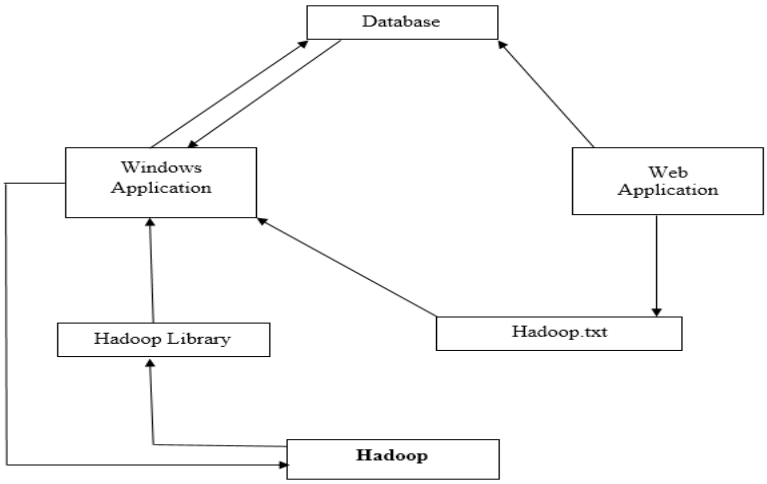


Figure 1: Architecture diagram of Working System (Dubey, 2016)

2.1.3. TECHNOLOGIES TO PROTECT PRIVACY

Governments and businesses are anxious to extract the valuable information which Big Data holds. When the thought of Big Data comes to mind, one thinks of the volume and variety of this data. But, issues such as the quality of the data, data privacy and security are of utmost importance and as Big Data gains momentum, these concerns will grow. Indeed, some experts believe that without the proper levels of security in the Big Data world, the consequences could negatively affect the development of this technology. (Thuraisingham, 2015)

Essentially, the elements of Big Data security consist of the following; infrastructure security, data privacy, data management, and integrity and reactive security. The International Organisation for Standardisation has embraced the classification of Big Data security into four areas, by creating a security model for Big Data security. (Moreno et al, 2016, P2)

INFRASTRUCTURE SECURITY

Before examining infrastructure security, it is important to mention that the discussion will involve the Hadoop technology, as this is the most commonly used data analytics software.

Some experts have proposed changing the Big Data architecture, so as to increase the security of the system. This could be achieved by developing a new architecture designed around the Hadoop system, in combination with network programming and muti-node reading. Furthermore, by developing new protocols and altering the configuration of the nodes, safe communications in vast networks controlled by Big Data could be realised. (Qin et al, 2016)

The authenticity of the Big Data, after it has been processed is important to ensure that it has real value. (Moreno et al, 2016, P5) Other experts have argued that the issue of authentication could be resolved by developing an identity-based signcryption system for Big Data. (Wei, 2016)
Signcryption is a public key cryptography system, which performs the tasks of digital signature and public key encryption concurrently. (Zheng, 1997)

DATA PRIVACY
Data privacy is an issue of great significance, for members of the public and also business, which avail of Big Data processing methods.
Consequently, numerous systems have been developed which aim to ensure that data privacy is provided.

Cryptography is such a method to ensure that the privacy of data is protected. Some academics suggest that cryptography can be used in Big Data privacy protection, by using a bitmap encryption system that can provide citizens' with their data privacy. (Yoon et al, 2015)
Others argue for processing data which is encrypted, and to analyse and control alterations with PigLatin with respect to encrypted data. (Stephen et al, 2014)

A traditional method to assist in protecting privacy is access control, which restricts access to those who are genuine system users. In the Big Data system, this could be achieved through a framework that integrates access management functions. (Colombo, 2015)

Some experts also examined how Map-Reduce could be used to protect privacy, and propose a system that can implement security policies at key-value level. (Ulusoy et al, 2015)

To further strengthen privacy protection, confidentiality systems can be employed, for instance by using masked data. This enhances data privacy by enabling specific computations to be performed on masked data. (Kepner et al, 2014) This can also be achieved by the Trusted Scheme for Hadoop Cluster (TSHC), it generates a new design system for Hadoop that increases the confidentiality and security of the data. (Quan, 2013)

Anonymising data can also be carried out to enable data privacy, this process involves using software, which can alter or eliminate confidential information from the data set. Furthermore, some academics recommend using a hybrid technique to anonymise the data. This involves combining the two most popular anonymisation systems, which are the Top Down Specialisation (TDS) and the Bottom Up Generalisation (BUG). (Irudayasamy et al, 2015)

DATA MANAGEMENT

Data Management is concerned with how to securely and safely manage the data after it has been collected and processed. In order to guarantee that the data is adequately protected once it has been collected, it is necessary to form a parameter that can determine an appropriate level of privacy. (Cheng et al, 2015) An additional method that could be used to protect data at the collection point involves separating the data and allocating data sets to specific Cloud storage service providers.

For businesses to realise the full value of their data, data needs to be shared amongst the particular data group that the Big Data is operating in, or alternatively, share the results among the participants. But, this practise raises concerns with regards to ensuring that there is a suitable standard of security and confidentiality when sharing this data between dissimilar parties.

These concerns could be eliminated by making the transmission of data secure, which can be done by employing a method founded on nested sparse sampling and co-prime sampling. (Chen et al, 2015)

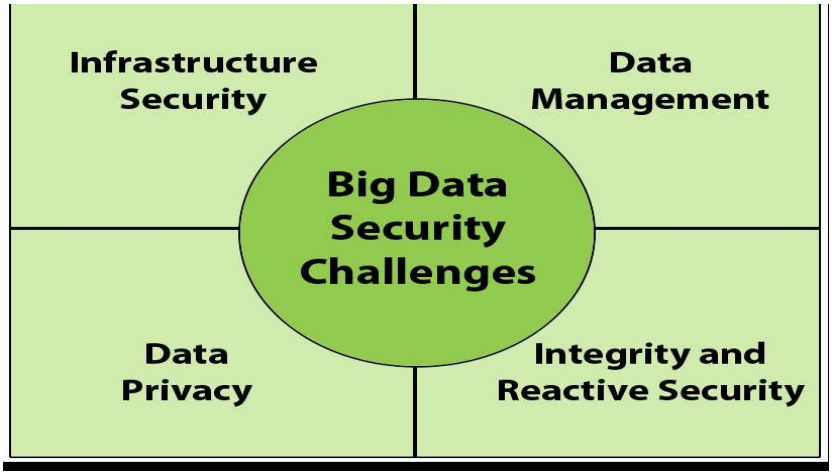


Figure 2: Main challenges as regards security in Big Data security (Moreno et al, 2016)

BIG DATA CONFIDENTIALITY IN DATA PRODUCTION STAGE

Data generation or production can occur by means of active production and passive production.

Active data production occurs due to the data owner actively and knowingly giving the data to a third party. Passive data production occurs because of the data owner's online activities, such as when they are browsing the Internet. In these circumstances the data owner may well not know that their data is being collected by a third party. (Mehmood et al, 2016, P3) Clearly the data owner wants to protect their important information and is also particularly anxious about the level of control that they may have over this data.

To decrease the possibility of a privacy infringement during the data generation stage, it is necessary to restrict access or falsify the data.

ACCESS RESTRICTION

When the data owner is providing their data passively, in other words when they are browsing the Internet, they can take precautionary steps so as to protect their privacy. They may utilise software tools like anti tracking extensions, script or advertisement blockers and encryption functionalities. These functionalities can reduce access to important and sensitive data. (Xu et al, 2014)

FALSIFYING DATA

On occasions, it is likely that it can be impossible to avoid unwanted access to sensitive data. In these situations, data can be altered using specific software tools prior to data being taken by a third party. The altering of the data makes it very difficult for a third party to access important and private information.

THERE ARE A NUMBER OF SOFTWARE TOOLS WHICH THE DATA OWNER CAN USE TO FABRICATE OR FALSIFY THE DATA;

The software tool socketpuppet can be used to conceal the online identity of the data owner. When numerous socketpuppets are used, the original data owner's data will be perceived to belong to many other people. Consequently, the third party data collector will be unable to connect the numerous socketpuppets to the original data owner. (Mehmood et al, 2016, P4)

The software tool Mask Me can assist the data owner in protecting their privacy, by masking their identity. It enables the data owner to create an assigned name or number for important information like their email address or their credit card number, which is obviously useful when shopping online for example. (Jain et al, 2016, P6)

BIG DATA PRIVACY AT THE DATA STORAGE STAGE

As a result of the reduction in storage costs and efficiencies in storage software, storing large amounts of data is not an issue. But, the challenge arises if the storage system is affected by a privacy breach, as it can result in individuals' sensitive data being divulged. (Sokolova, 2015)

PRIVACY PRESERVATION FOR CLOUD STORAGE

There are three important elements to data security in the Cloud, these are confidentiality, integrity and availability. (Xiao et al, 2013) Confidentiality and integrity are both linked to the privacy of the data, as if one or both of these elements are infringed, this will clearly have a harmful outcome for the data subject's privacy.

Availability of information essentially means that those who are permitted, should be permitted to access the data if they require it.

There are a number of protection methods which can be employed to help guard a citizen's privacy in the Cloud storage system, for instance a data owner can encrypt their data by utilising public key encryption. Only the designated recipient may decrypt the data originally sent. The following are the other protective methods:

ATTRIBUTE BASED ENCRYPTION

The data owner determines what the access conditions are, while the data are encrypted in accordance with those guidelines. Decryption of the data can only be allowed by the data owner whose attributes match the access guidelines as set out by the data owner. (Bethencourt et al, 2007)

HOMOMORPHIC ENCRYPTION

Homomorphic encryption enables computations to be carried out over encrypted data stored in the Cloud. (Gentry, 2009)

STORAGE PATH ENCRYPTION

Storage path encryption service provides a safe storage of Big Data in the Cloud, when in the Cloud, Big Data is split into public data and confidential data. The storage path to the Big Data store is encrypted, as opposed to the Big Data being encrypted on mass. This encrypted path is named the cryptographic virtual mapping of Big Data. (Hongbing et al, 2015)

USE OF HYBRID CLOUDS

The hybrid Cloud service enables the operation of a combination of on site, private and third party Cloud and public Cloud services with management among the two specific platforms. (Jain et al, 2016, P7)

PRIVACY PRESERVING METHODS IN BIG DATA

There are a number of privacy preserving methods that can be used to protect Big Data, such as anonymisation and cryptography.

Anonymisation is the traditional form of privacy protection and it involves private data being altered in a manner that makes it almost impossible for the data subject to be re-identified. Also, the data that relates to them cannot be revealed. (Parise, 2011)
But, because Big Data consists of vast amounts of data and ever increasing availability of analytics software, the current anonymisation methods no longer provide the adequate levels of effective protection.

Furthermore, these anonymisation methods are designed to anonymise static data, while Big Data sets are of a dynamic form. (Enisa, 2015)

ENCRYPTION
Encryption is an important privacy preserving method, it modifies the data so that only the persons who have permissions may examine it. The data is modified by using encryption algorithms and encryption keys, which must be secure. (Cavoukian, 2009)

IDENTITY BASED ENCRYPTION
Identity based encryption is designed along the lines of public key encryption, but the key management system is simplified. This is because it uses the public key technology, by making use of the data subject's identities, such as their email address or their Internet Protocol address as public keys. This encryption method ensures that the anonymity of both the recipient and sender are secure. (Boyen, 2006)

2.1.4. ARE THESE TECHNOLOGIES GOOD ENOUGH TO EFFICIENTLY PROTECT PRIVACY?

THE LIMITATIONS OF THE HADOOP SYSTEM SOFTWARE.
The Hadoop system is not very efficient, because when it is analysing and processing data it uses a parallel processing system to process data, which results in duplication of data.
As mentioned earlier, the Hadoop system is an open source software system, but this system can result in a likely difference in quality. The open development process leads to a situation where there is no incentive to provide software quality. This can be the case, as within a development team, there can be a level of competition and diverse interests. (Conboy, 2014) Under the open development process, there are many factors that can

affect quality management. For example, the development methodology is frequently undocumented, testing and quality assurance are methods are informally applied and only a small number of measurable quality goals are defined. (Aberdour, 2007) Studies have also revealed that the initial version of open source software projects have higher defects. (Javed et al, 2016) Furthermore, the open source development system is a way for budding software developers to engage in the process of software programming. In saying this, there is a section of experienced programmers who assist in the development of open source assignments, but they assist as a result of their interest in the particular development plans. (Paraccel, 2012, P4)

The Hadoop Map-Reduce system operates as a file system on clusters of a random size, hence it was deemed reasonable to not include an efficient storage system in the design. The disadvantage of the HDFS system is that there is no optimiser, in effect the developers will have to ensure that they optimise their data stream. Furthermore, because it is based on a file system, it is not possible to have recovery checkpoints or data management consistency. Consequently, the results which are drawn from a Hadoop cluster may not always be accurate.

The Hadoop system is powerful and effective for most data management issues, but it needs to be operated by an expert in order to achieve the benefits of the system. (Paraccel, 2012, P5)

THE SPECIFIC LIMITATIONS OF HADOOP

TOO MUCH DUPLICATION OF BIG DATA

The HDFS file system is not very efficient, and consequently three copies of data are produced. There are three copies initially and due to the requirement for data to be local to ensure high performance, six copies of data may be produced.

ALMOST NONEXISTENT SQL SUPPORT

Within the Hadoop framework there is function that enables the use of queries. However, they are at a primitive level, as sub queries or group by function do not exist.

DATA EXECUTION THAT WASTES RESOURCES

The HDFS file system does not have a query optimiser. As a result, it is ineffective at managing data execution and the Hadoop data clusters are bigger than the usual size for other databases.

REQUIREMENT FOR SPECIFIC SKILLS TO USE HADOOP.

In order to use the Mahout data mining libraries, the operator needs to have some expertise of the algorithms. (Paraccel, 2012)

PRIVACY PRESERVING DATA MINING

As Big Data continues to grow, so too does the importance of ensuring that the data is secure, when it is transferred through the Internet. Privacy preserving data mining provides a mechanism whereby this can be achieved, due to the complex data mining algorithm.

There are a number of privacy preserving methods which are used to carry out the data mining process. The following are the most common methods; clustering, *K* anonymity, classification, distributed privacy preservation, association rule, randomisation, cryptographic, *L*-diverse, condensation and finally taxonomy tree. (Sachan et al, 2013)

These methods protect the data by altering it, which results in concealing or deleting the important and sensitive data to be masked. The methods provide the means to resolve the data subject's original data from the altered data. (Xu et al, 2011)

The privacy preserving methods also can operate, by utilising data distribution and dispersed partitioning throughout numerous units.

PRIVACY PRESERVING DATA MINING METHOD ALGORITHMS

DATA DISTRIBUTION

Privacy protection data mining is performed on distributed data by an algorithm.

The distributed data is partitioned both vertically and horizontally.

DATA DISTORTION

The data distortion method algorithm firstly changes the data base record, then the algorithm makes amendments to the attribute value of the data.

DATA OR RULES HIDDEN

This method's algorithm conceals the original data or the original data rules. (Zong et al, 2012)

FLAWS OF PPDM METHODS

There are a number of performance issues in respect of the current privacy preserving data mining methods, which show that they are not very effective. Some of these issues are the effectiveness of data, scalability, the reliability of the data mining and the overhead performance. (Aldeen et al, 2015, P17)

In order to overcome these issues, a reliable, scalable and effective system of methods needs to be developed. Furthermore, there should not be a relationship amongst personal data and the personal identification number or code.

While K-anonymity is considered to be a reliable privacy protection method, some experts established that the data managed by the method, were vulnerable to attacks and prone to Internet phishing. (Aldeen et al, 2015, P18)

K-anonymity therefore has to be redesigned with a superior data infrastructure, to be able provide the range of tasks. The current seeking algorithms are fast when retrieving data, but they cannot scale up to a bigger quantity of data. This is as a result of the linear increase in response time, relative to the quantity of the explored datasets. (Aldeen et al, 2015. P18)

Big Data analytics, with its ability to plan and predict for the future, is seen as an important tool in health, science and even astronomy. It is becoming more common for third parties to carry out Big Data processing on private data, which obviously raises an issue of breach of the data subjects' privacy.

To prevent such a breach, the algorithms such as the association rule mining, should be organised and designed to ensure that privacy is preserved. There may be instances where the data which is retrieved from a business, does not contain any meaningful or valuable information. Furthermore, obtaining the data may be problematic, because it may not be possible legally or because of the potential privacy infringement.

These issues could be resolved if privacy preserving and distributed analytic frameworks were properly designed. Specifically these systems would be capable of processing various datasets from associated business, and simultaneously ensuring that each dataset's privacy is protected. (A. Mehmood et al, 2016, P1832)

Some academics suggest that the use of homomorphic encryption, which is a secure multiparty software tool, could be used to manage such difficult issues. But, the use of this encryption method in respect of Big Data analytics, can mean that there will be overly complex data processing. (A. Mehmood et al, 2016, P1832)

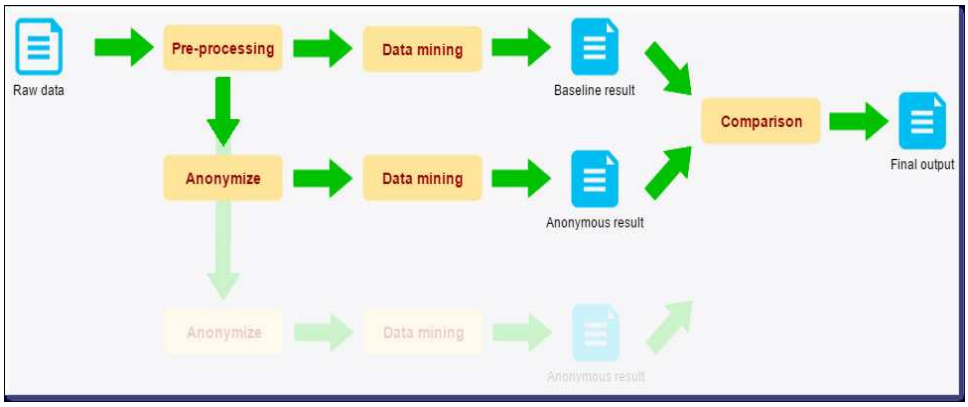


Figure 3: Representing the overall data mining process (Willemsen, 2016)

THE LIMITATIONS OF ANONYMISATION

In a number of ways the anonymisation privacy methods are less than perfect. The following are the issues that highlight the limitations.

DATA SUBJECTS LACK OF CONTROL OVER THEIR DATA

The anonymisation process at present does not allow the data subject to have control over their information and they cannot ensure that their data is properly protected. The data controller has a leading role to play with respect to the data, as he or she is legally in a position to manage this data by choosing the anonymisation method, the parameters and the privacy standard.

ADVERSARY'S BACKGROUND

The K-anonymity process consists of the utility-first method and the privacy first method. As part of these methods, restrictive assumptions have to be made with respect to what the data hacker or thief's possible awareness is. In effect, it is reasonable to expect that the hacker can connect with peripheral datasets by means of separate quasi-identifier attributes. But, with e-differential privacy, such assumptions do not need to be made. As when there is a deviation in the system, this makes the presence or absence of a specific data file unrecognisable in the anonymised data. Therefore, the usefulness of the data is severely restricted.

2.1.5. SUMMARY

The era of Big Data has resulted in major changes and disruption in the manner in which our personal data is managed and processed. The threats to our data are as a result of unauthorised disclosure, loss and theft, the outsourcing of data analytics and the secondary use of data. The most important element in all of this is data privacy, hence the need to have some sort of discussion on the Legal and IT side of privacy

The possibility of achieving privacy in the Big Data world can be a reality, including making use of data analytics to help organisations to become more innovative. To prevent the cat getting out of the bag, it is necessary for organisations and individuals to ensure that privacy protections are in place at the outset.

Apart from the Hadoop technologies and other privacy measures that were examined in this section, privacy by design, data minimisation, differential privacy, de-identification and synthetic data are others ways to protect data in the Big Data age and into the future.

2.2. LEGAL SIDE OF BIG DATA AND PRIVACY RESEARCH

2.2.1. INTRODUCTION

The Big Data era has arrived and will continue to be a dominant force that will affect not just the technological and data security sectors, but also government, science and business communities. Big Data consists of volumes of structured and unstructured data. In this ever evolving technological world, most if not all organisations, whether they are governmental,

a company or nongovernmental, collect and store personal data about their customers, citizens or employees. (Custers, 2016, P1)

The power of Big Data Analytics, which is the algorithmic analysis of datasets (Ulbricht, 2016), means that relationships, trends and sequences can be found by merging vast quantities of data from various sources. (Custers, 2016, P2) Due to the analytics processing power, and resulting analysis, governments can make more informed and tailored policy decisions and scientists can make new life changing discoveries. (Oostveen, 2016) This Chapter, which is the Legal section of the literature review, examines the important legal principles and issues that will have a bearing on Big Data analytics currently and in the immediate future.

The first part of the Chapter briefly discusses whether Big Data will have an effect on the legal process with respect to whether the appropriate regulations will be implemented to manage the Big Data world. Next, there is a brief layout of risk mitigation schemes, which can help social media users and others to protect their personal data.

The second part of the Chapter goes into detail regarding the reality of data protection law at present, and how the law is managing to keep abreast of the technological developments. This includes a discussion of the purpose limitation principle, which is part of the GDPR, where data can be collected if there is a specific and legitimate purpose. Then the data minimisation principle is analysed. This principle stipulates that companies must reduce the quantity of data that is collected and processed, and indeed not collect any more than is required in order to realise the exact aim of the collected data. But, firstly, the concept of consent is examined, this involves viewing the relationship between consent and the new GDPR and the complexities involved.

2.2.2. IS BIG DATA GOING TO SHAKE UP THE LAW?

The widespread collection and reuse of personal data and other data, at unprecedented levels, has prompted the discussion of protecting individuals' data protection and privacy rights. This is especially the case given that data analytics has lead to wide scale electronic surveillance, profiling, and leaking of private data. In order to balance the competing interests of citizens anxious to protect their privacy, and companies and government, who

are keen to extract the unique power of Big Data analytics, it is essential to implement regulations that can manage the process for both parties.

RISK MITIGATION SCHEMES

There are a number of ways in which the risks to personal data privacy as a result of Big Data analytics can be reduced, they are outlined briefly below.

AWARENESS

If individuals are made aware of their privacy rights through a point of contact, this will assist greatly in their ability to exercise their privacy rights.

USER CONTROL

In order to enable the individual to have more control over their data online, the availability of privacy preferences and personal data stores could help them in their use of online applications.

RETENTION, DELETION AND ANONYMISATION

In a situation where the citizen's data is stored as a result of a smart parking application for example, their data should be deleted when they have parked their car, unless they agreed for their data to be retained. If it is necessary for their data to be retained for a longer period of time, then the data should be anonymised. (Enisa, 2015)

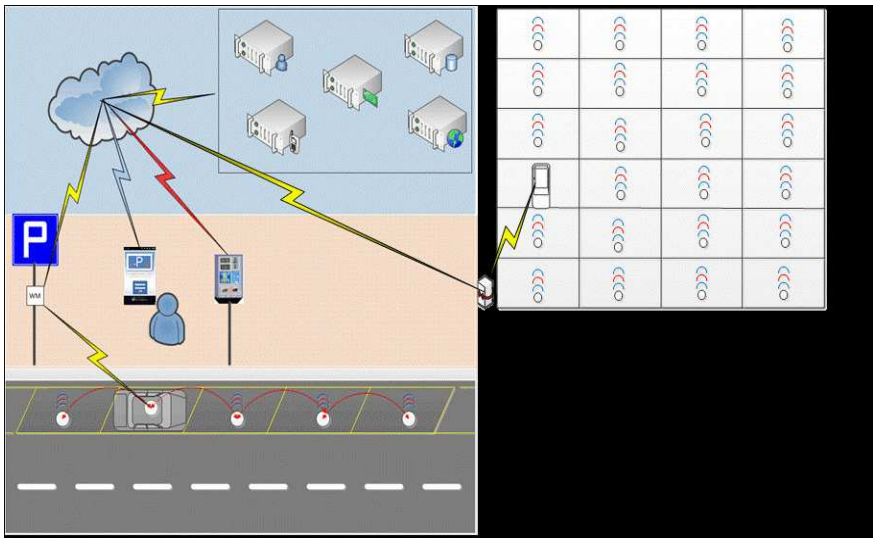


Figure 4: Smart parking application scenario. (Enisa, 2015)

THE STORY SO FAR, DATA PROTECTION LAW

The European privacy and data protection legal framework is based upon fundamental rights. (Chen, 2016, P310) This is rather different from the American system of data protection, which typically has a sector specific method of legislating data protection and privacy law. Also these laws are implemented at a state level as opposed to federally in the United States.

The building block, on which data protection law in Europe is based, is the European Union (EU) Data Protection Directive. (Directive 95/46/EC, 1995) This twenty year old Directive is being replaced by Regulation 2016/679, or the more commonly know, General Data Protection Regulation. (General Data Protection Regulation (Regulation (EU) 2016/679, 2016). This regulation comes into effect in thirty one European Union and European Economic area countries, from the 25th of May 2018. (Chen, 2016, P310)

The implementation of a Directive requires member states to transpose it into national law, through an act of the respective national parliament. The implementation process involved for a regulation is less complicated, as it does not need to be transposed into national law. The regulation has direct effect in all member states, and this ensures that there is a corresponding set of core rules.

Data protection and privacy law is based upon fundamental rights, which are protected and enshrined in treaties such as the European Convention on Human Rights (ECHR) and the Charter of Fundamental Rights of the European Union. (Broeders, 2017, P316). Specifically, data protection is protected by Article 8 of the EU Charter of Fundamental Rights, and a distinction has to be made between this right and respect for private and family life, which is contained in Article 7 of the Charter.

2.2.3. THE LAW AIMING TO KEEP PACE AND CONTROL THE POWER OF TECHNOLOGICAL ADVANCEMENT.

There is a balancing act required by lawmakers so as to ensure that the positive effects of Big Data are not hindered by the desire to ensure citizens' data protection rights are protected. Big Data processes involve collecting and working with aggregated and unfocused data, general patterns and group profiles, and to a lesser extent involve the processing of data at a personal level. This data is subsequently linked and reused for other purposes by other parties, such as social media companies or online advertisers such as Google. Big Data processes promises to provide more accurate and effective data. This can prove to be fruitful to the business sector, as they can be aware of their customers' preferences through cookies and therefore can ensure that their marketing is more targeted and effective. An example of this can be seen when an individual is using a news web site or similar, and at the bottom or along the sides of the site, there are advertisements that are associated with an earlier search of product by the individual.

It seems clear that presently, the European regulatory framework is designed to principally deal with data collection, which is the first stage in the Big Data process.
Furthermore, this framework consists of legal values which are at times in conflict with the Big Data analysis model. (Broeders, 2017, P317)

Indeed, the Big Data process is creating a strain on legal principles, such as the purpose limitation and data minimisation. These principles are of critical importance to the data collection stage of the Big Data process. (Broeders, 2017, P317) Large volumes of data are collected as part of the Big Data analytics process, indeed mainly more data than necessary is collected, and the benefits of Big Data analysis stems from the use of secondary data. Therefore, the conflict arises because the data minimisation principle

requires that data should be collected where it is necessary for a particular objective and the purpose limitation principle specifies that data may only be gathered and stored where there is a specific reason for the processing operation. These legal principles will be examined in the next section below.

CONCEPT OF CONSENT

Consent is an important element in data protection law. In particular individual consent is essential, to provide a legal basis before personal data can be processed.

The Data Protection Directive has provided data subjects with a strong level of data protection, but as the Internet technology advances, the level of legal protection needs to be stronger.

The GDPR takes a stricter approach with respect to attaining the data subject's consent. This is because there must be specific consent for a specific processing action, the exclusion of consent in the terms and conditions and the data subject's explicit right to remove their consent without restriction. (Mason Hayes and Curran, 2017, P14)

The new regulation requires that the data subject's consent must be freely provided, specific, unambiguous and informed, before a company can lawfully process the personal data of an individual.

Essentially, this means that the data subject has to know what the company intends to do with their personal data, also there must a specific indication that the individual consents.

But, to ensure that the consent is legitimate, a further four conditions must be fulfilled. (Mason Hayes and Curran, 2017, P14)

Firstly, there is an onus of proof on the data controller, that the data subject has given their consent for the data processing action.

There must be an independent consent clause, thus if consent is provided in a written contract, which can consist of other issues, the consent request must be clearly evident in the contract. (Mason Hayes and Curran, 2017, P14)

The consent must be voluntary. In other words, if the consent is freely provided, the effecting of a contract should not be restricted based on the citizen consenting to the processing of their data, which is not required for the effecting of the contract. (Mason Hayes and Curran, 2017, P15)

Some experts argue that the notice and consent mechanism, in which a company informs individuals about what will be done with their data, is not workable in the Big Data age. It seems to be very difficult to establish that the data subject has provided unequivocal consent, considering that the data analysis using artificial techniques is obscure to say the least. (Buttarelli, 2016)

The issue of consent in the era of Big Data analytics is complicated indeed. The new GDPR states that when the data controller is relying on the consent of an individual, before their data is collected, the process for determining consent must ensure that consent is:

"unambiguous", in the data that is not sensitive (Article 6 of the GDPR which states that consent is required, and Article 4 which identifies consent to be "unambiguous"); (Article 6 and 4 GDPR, 2016)

"explicit", when the data are sensitive data (i.e. relevant to any of the types of sensitive data listed in Article (Art) 9(1) of the GDPR, for example physical or mental health data.). (Article 9(1) GDPR, 2016)

The question could be asked as to how the data controller can receive appropriate consent. The GDPR declares that consent, if it is explicit or not, must be provided "by a statement or by a clear affirmation action" which signifies conformity to the processing of personal data" (Art 4). The recital 32 elaborates on this when it says that consent "should be given by a clear affirmative act, such as a written statement, including by electronic means, or an oral statement."

The GDPR also says that "Silence, pre-ticked boxes or inactivity should not constitute consent" (Recital 32). Recital 32 clarifies further when it states that consent may be provided through "another statement or conduct which clearly indicates in this context the

data subject's acceptance of the proposed processing of his or her personal data". (Recital 32 of GDPR, 2016)

Effectively, to provide explicit consent there will have to be an affirmative action on the part of the individual, for example, ticking the box, and to satisfy the requirements for unambiguous consent and an affirmative action will be sufficient. (Scaife, 2016)

Article 7 of the GDPR also ascertains that the burden of proof on existence of individual consent for a particular processing operation lies with the data controller. (Scaife, 2016)

It is clear that the regulation takes individual consent seriously, as a result of the provisions that are included. Most of the current data processing of individuals' data, which is carried out by data controllers, is based on it. The indirect and explicit ways of determining consent cause some difficulties. The initial commission proposal was anxious to protect individual rights and wanted there to be explicit consent in the circumstances. However, this was rejected by the Parliament and the Council, which was somewhat understandable, because of the complexity of current processing. De Hert states that those strict conditions that were included in the final text of the regulation, should remain in place, to ensure that there is the strongest level of data protection for citizens. (de Hert et al, 2016)

PURPOSE LIMITATION PRINCIPLE

The purpose limitation principle (Article 5(1)(b) GDPR,2016 and Article 6(1)(b) DPD, 1995) requires that data can be gathered if there is a specific and legitimate purpose, what is known as the original purpose and the data may not be processed in a manner contrary to those original purposes. (Broeders, 2017, P317)

Article 6(4) (c) GDPR provides a more detailed explanation of the principle and states that (where consent of a Member State law does not apply) when determining if a purpose matches up with the original purpose for the collected data, the data controller must consider, "the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10". (Article 6(4) (c) GDPR, 2016)

Consequently, an awareness of the nature of personal data, and which groups of data are processed, can assist in evaluating if a purpose matches up with the original purpose for collecting the data. (Cradock, 2017, P144)

Some academics believe that the purpose limitation principle creates difficulties in the development of Big Data analytics. (World Economic Forum, 2013) This is because the purpose limitation principle negatively affects the notice and consent model. For example, Big Data analytics facilitate data analysis, which uses different algorithms that reveal unanticipated associations that can be utilised for new purposes. Consequently, the purpose limitation principle limits a company's ability and autonomy in their wish to be innovative. (Ghani, 2016, P119) This also highlights an opinion of Big Data analytics as a constantly changing process, the analysis of data using different algorithms, resulting in unexpected relationships, which can cause data to be used for new purposes. Such as purposes, that the data subject may not have agreed to or be aware of. (Information Commissioner's Office, 2016)

Some other experts argue that the purpose limitation principle reduces a company's autonomy in being innovative and inventive. The essential point in respect of the principle is that it thwarts random data reuse, but it does not need to be a complete obstruction to extorting value from data. The question that has to be posed is how can compatibility be determined. This is the prevailing view of most academics with respect to the purpose limitation principle. (Information Commissioner's Office, 2016)
The United Kingdom (UK) Information Commissioner's view is that a vital point when deciding if a new purpose is mismatched with the original purpose is if it is fair. (Information Commissioner's Office, 2016) Fundamentally, it is essential for the relevant company to take into account how or if the new purpose can impinge on the privacy of a citizen, and if it is likely that the individual would expect that their data could be used in this manner. (Information Commissioner's Office, 2016) The following three paragraphs discuss the importance of helping individuals understand how their personal data is processed and the benefits for the data controller and data subject if the data is divided into categories.

In order to help citizens understand how their data is processed, it may be necessary to divide personal data into categories. Thereby, creating an environment where the privacy policies could be shorter in length, because the usual amount of information required to inform the data subject could be reduced. (Cradock, 2017, P144)

The categorisation of personal data can ensure that there is an evaluation of risk related to the processing of the data. Furthermore, it can highlight which technological and organisational methods are required, and assist on deciding if a secondary purpose or use of data is compatible. (Cradock, 2017)

In a scenario in which different rights or requirements pertain to particular categories, it enables the data owner to consider if the data controller has fulfilled his or her obligations, when the data owner is made aware of the categories of data processed.
Indeed, the law needs to provide clarity on when data controllers should notify individuals of the categories of data they process in respect of the requirement to inform. (Cradock, 2017, P145)

DATA MINIMISATION PRINCIPLE

The data minimisation principle is a fundamental component of data protection law. The data minimisation principle stipulates that companies must reduce the quantity of data that is collected and processed, and indeed not collect any more than is required in order to realise the exact aim of the collected data. (Broeders, 2017, P318) Additionally, the collected data should be removed, as soon as the aim has been realised. (Broeders, 2017, P318)

The GDPR includes a section on the data minimisation principle, where it states that; "personal data shall be "adequate, relevant and limited to what is required in relation to the purposes for which they are processed". (GDPR Article 5(1)(c), 2016)

The business practice of collecting vast amounts of unnecessary data, is putting data protection at risk, this is particularly the case with respect to linked location data. Linked location data can reveal an individual's specific actions over a set amount of time and

consequently enable the discovery of the particulars of the individual's private life. On the other hand, if data was collected when it was required for a specific purpose, this could ensure that businesses do not have any difficulties in finding the necessary data.

Principle five of the Data Protection Act in the UK states that data processed for any function or functions shall not be kept for longer than is required for that function or those functions. (Data Protection Act, 1998)

But, in the age of Big Data analytics, putting the intentions of principle five into practise may prove to be difficult. Especially, as the ability to store data continues to rise and the associated costs are decreasing. Also, the power of analytics to process large amounts of data may prompt data controllers to store the data for excessively long periods of time. (Information Commissioner's Office, 2016, P37)

Additionally, the Article 29 Working Party have stated that the continuation of the purpose limitation principle is crucial, to guarantee that businesses which hold a dominant position, before the advancement of Big Data technologies, do not have an overly advantageous position over new entrants to the market. (Europa.eu, 2014) The Working Party was established under Article 29 of Directive 95/46/EC and it involves a spokesperson from the data protection authority of all EU Member State, the European Data Protection Supervisor and the EU Commission. The Working Party aim is to synchronize the use of data protection law in all the member states, and distributes opinions and recommendations on a number of data protection issues. (dataprotection.ie, 2017)

Perhaps, companies should be obligated to clarify at the beginning, the reason for collecting and processing specific datasets. Also, they should express clearly what discoveries they believe will be made or what they can achieve by processing the data. Thereby, establishing that the data is pertinent and proportionate, in respect of the aim. (Information Commissioner's Office, 2016, P41)

2.2.4. SUMMARY

The growth of Big Data analytics and the processing of personal data at an ever increasing pace, means that the law has to work in conjunction with this technology, in order to

protect individuals' personal data. That's why the legal section of the literature review consists of an examination of the important legal principles that will try to control the processing activities of Big Data analytics.

The first part of the Chapter discussed some risk mitigation schemes, which could be used by citizens, to reduce or minimise the risks to their personal data privacy, as a result of Big Data analytics. These schemes included user control, whereby the privacy preferences could help individuals to protect their privacy online, and awareness, which involves a point of contact to assist in the protection of users' privacy rights.

Next the concept of consent is examined, this involved analysing the relationship between consent and the new GDPR and the complexities involved therein. The new regulation takes a strict approach with respect to the standard of consent provided by an individual, before an organisation can use their data. In particular, there must be specific consent for a specific processing action and that the data subject's consent must be freely provided, specific unambiguous and informed, before a company can lawfully process the personal data of an individual.

Following this there was an examination of the purpose limitation principle, which originates from the GDPR. This principle aims to protect personal data by ensuring that data can only be collected by an organisation, if it has a specific purpose for doing so. In effect the principle tries to reduce random and unwarranted data use and this helps to increase the likelihood of personal data being protected.

Finally, the data minimisation principle was discussed. This principle demands that companies must reduce the quantity of data that is collected and processed, and not collect any more than is required in order to realise the exact aim of the collected data.
The intention of this principle is to reduce the practice of collecting vast amounts of unnecessary data by organisations, and thus ensure that individuals' data protection is not put at risk by this practise.

The GDPR will ensure that the companies do not have it all their own way with respect to using individuals' data as they wish.

There must be legally compliant security and data protection by design, which is required under the GDPR. The answer going forward is to have controlled linkable data, which entails restricting secure access to the data, which is specifically required for the data analytics process at a time.

In order to protect the rights of personal data, data protection law exists to do so, but the GDPR ensures that the levels of protection are even higher. For instance, there is greater responsibility on data controllers to show that they are analysing and processing the data properly and lawfully. The companies who carry out data processing and analytics activities, must ensure that they legally compliant or else they have to pay significant fines as part of the new guidelines under the GDPR.

3. CHAPTER THREE – RESEARCH METHODOLOGY

3.1. QUANTITATIVE RESEARCH METHODS

3.1.1. INFERENTIAL STATISTICS

The Hypothesis test is a formal way to examine a null hypothesis, which is a statement about population parameters. (Epstein and Martin, 2014, p155)

In order to think about a null hypothesis from a legal perspective, it is worth thinking about a criminal trial. The trial starts with a null hypothesis that the accused is not guilty. (Epstein and Martin, 2014, p155) The jury in the trial then look at the evidence, to see whether it is consistent with the null hypothesis that the accused is not guilty or whether it is consistent with the alternative hypothesis, the accused is guilty. (Epstein and Martin, 2014, P155).

Inferential statistics make an inference about data, as they are samples. As researchers, we must ask ourselves, how confident are we about making inference?

In a legal setting the example of inferential statistics, can be seen in the Griffin legal. In Grifffin (Griffin v. Board of Regents, 1986) the district court ruled that an inference of discrimination could not be found, where there was an R-Square of .45, as the level of determination was too low. (Luna, 2006, p212)

It is argued by Ballinger that inferential statistics are unsuitable for development studies, when being used in respect of superpopulations. This is when data is calculated by an actual world sample or an evident population and not the product of a random example from a bigger population. (Ballinger C, 2011)

The student's research title 'Big Data Analytics: Balancing Individuals' Privacy Rights and Business Interests', is in itself a research question that is a hypothesis. The hypothesis in this case is that the privacy and data protection rights of individuals are being infringed, because of the way that companies use this data for their financial benefit.

This hypothesis will be accepted or rejected based on data or findings that will be collected as part of the research process. The data will be produced by surveys, which will assist the author in making an inference about the data. In order to test the hypothesis, a comparison will be made between those Internet users who believe that their privacy has been infringed and those who feel that it has not been infringed.

3.1.2. NON-PARAMETRIC TESTS

Non-Parametric tests work on discrete categories and are superior in particular situations, such as in small sample settings. In effect, non-parametric tests can be described as statistical tools that measure frequencies or ranks, which are not entirely quantitative in nature and may in fact be qualitative, with respect to what is analysed. In a legal context, an example is the frequency with which a judge votes in a particular way on a particular policy issue. (Boyd, 1972, p291)

Non-parametric tests are not as constrained with respect to the assumptions that must be met in applying them. They are also less dependable in relation to the inferences to be garnered from the outcomes of their usage.

The research paper, 'Unintended and Persistent Consequences of Regulation: The Case of Cable Television Provision in Canada', highlights the benefits of using the research method; non-parametric tests.

The journal paper's authors illustrate their use of the non-parametric scale measurement techniques, to gauge the effect of regulation on cable providers in Canada. (Law & Nolan, 2003, p395) The authors had to use the non-parametric scale measurement techniques, because the smaller size of the parametric sub-samples provided an inadequate inference from the parametric model. (Law & Nolan, 2003, p395)

But Dallal points out that non parametric waste information. (Dallal 2014) For instance, the sign test uses only the signs of the observations, and ranks safeguard information about the order of the data but remove the actual values. Due to the fact that information is redundant, nonparametric tests are not as powerful as the parametric tests.

The student's thesis will not be using non-parametric tests, because the data is not ranked and it is not measured in terms of frequencies. Also, observations of the data cannot be made from a normally distributed population, as the data is uneven.

3.1.3. PARAMETRIC TESTS

The parametric test is a statistical test that makes assumptions with regard to the parameters of the population distribution, and the researcher's data is extracted as a result.

The use of parametric tests in a legal or social science research setting can be seen in the research paper, 'If It Can't Be Lake Woebegone... A Nationwide Survey of Law School Grading and Grade Normalization Practices', by Robert C. Downs and Nancy Levit.

Part of their research paper involved the use of statistical analysis in relation to two sample groups of law students.

The authors used parametric tests, to show how grade variations between the student groups could be tested, in order to establish if the difference may be due to random chance or for some other reason, such as varying grading scales of the law professors. (Downs & Levit, 1997, p832)

While Downs and Levit's example highlights the advantages of the parametric tests method, the experience of Ostrovskii, illustrates a disadvantage with this type of testing. In particular, the z test, which is one of the testing facilitates in the parametric method, was not very good at measuring a precise single distribution. It is too sensitive to minor differences in the input data. (Ostrovskii, 2013, P18)

For the student's academic thesis, it will not be necessary to use parametric testing, as the research will not involve the use of a large volume of statistics. As the sample size of the student's data is too small for the benefits of parametric tests to be realised. In addition, there will be a stronger emphasis on qualitative research methods, such as interviews, as opposed to quantitative research methods.

3.1.4. MULTIVARIATE TECHNIQUES

It is generally accepted that quantitative research methods form the basis of scientific researchers' work, and have done so over many years. In recent times, it is evidentially clear that social scientists are more frequently using numerical analysis techniques, as a vital cog in their research process.

The increasing use of statistical techniques by social scientists, is because they enable social and economic relationships to be examined with the same empirical standards, which are associated with the laboratory sciences. (Arnold Lozowick et al, 1968, p1641) Quantitative multivariate analysis is such a technique, that the social scientist has now become accustomed to using. (Arnold Lozowick et al, 1968, p1641)

Effectively, this technique can turn a complicated legal question into a truthful answer, by using numerical analysis. (Arnold Lozowick et al, 1968, p1642)

The merit of multivariate techniques is evident in the Lozowick example, but in another case, there is evidence of its limitations. As (Savitri Abeyasekera, 2005) points out the cluster analysis, which is a multivariate technique, can be problematic when identifying an appropriate similarity or distance measure and with deciding which clustering method to employ.

A number of issues had to be considered, such as the data type and the power of the clustering technique, when managing small changes in the data.

The student's thesis will not use multivariate techniques as part of the research process for the academic thesis, as the size and complexity of the surveys does not require their use. The number of variables and sample size is too small

3.2. QUALITATIVE RESEARCH METHODS

3.2.1. QUESTIONNAIRES

The most common and popular quantitative data method is the Questionnaire, while it is also seen as a qualitative method. The questionnaire normally consists of closed questions, which result in simple statistical summaries, or open questions, which allow for a qualitative, lengthy and individual or specific response. (Hutchinson, 2006)

The academic paper by Lee Jarvis and Stuart Macdonald is a good example of the use of a questionnaire in the legal sphere and where the benefits of this research method are evident. In the paper, 'What Is Cyberterrorism? Findings from a Survey of Researchers', the authors published the results of a survey. (Jarvis and Macdonald, 2014, p657) The intention of the survey was to discover what the international research community thought about cyberterrorism. (Jarvis and Macdonald, 2014) Specifically, it explored the various views of its 118 respondents, on the importance of the requirement for an explicit definition of cyber-terrorism for legislators. (Jarvis and Macdonald, 2014, p657) Finally,

most of the researchers accepted that a precise definition of cyberterrorism is essential for policymakers. (Jarvis and Macdonald, 2014, p657) Questionnaires are not without their faults, as Beiske argues that questionnaires that are incorrectly completed will result in a reduction in the level or standard of data acquired. (Beiske, B, 2002)

The relationship between the social media users' privacy rights and the commercial goals of the social media companies was central in determining what shape the student's survey will take. The population that the questionnaire was concerned with receiving responses from, was those who use the Internet on a regular basis, such as those who access social media websites and some experts in the area of IT law. Variables were introduced to the questionnaire with the intention of revealing the truth. Variables such as the privacy concerns of Internet users and what measures can be taken to improve privacy law online.

The survey used a collection of open-ended and closed questions.
There were a number of closed questions, which were designed to produce quantitative data, and there were some open-ended questions that created qualitative data. The survey consisted of between ten and eighteen questions.

The open ended survey questions were based upon issues such as; development of an appropriate legal framework, which will resolve the issues with regard to privacy in the new world of Big Data.

The other issues that will be directed to the Internet users, will be concerned with whether they feel safe online and what can be done to help ease their fears online.

To ensure that there is a high response and completion rate, the questionnaire will be made available by an online survey, such as survey monkey.

An example of one of the questions that will be part of the survey will be as follows;

Question 1: On a scale of 1 to 10, where 1 is no, not at all and 10 is yes, definitive yes,
Do you feel safe when you access your social media account?

Question 2: If you answered yes, why do you feel safe online?

Question 3: If you answered no, why do you feel unsafe online?

3.2.2. FOCUS GROUPS

Focus groups are an obtrusive method of data collection, which involves respondents being part of a group interview. The researcher can benefit from the communication between the members of the focus group, in order to produce research data.

Emily Finch and Vanessa E. Munro's research paper, 'Lifting the Veil: The Use of Focus Groups and Trial Simulations in Legal Research', illustrates the use of focus groups in legal academic research. Their paper analysed the attitude of jurors in rape cases, which involved an intoxicated plaintiff. (Finch & Munro, 2008, p30)

Finch and Munro believe that research methods such as focus groups and interviews may well create significant benefits to the legal research sector. (Finch & Munro, 2008, p32) Although these research methods have been used over a considerable period of time, in the social sciences research area, the methods are not yet considered to be as important as other research methods, in the legal research community. The reason for this is that the law remains unconvinced about empirical methodology.

But, in a study by Smithson, she found that the use of focus groups has some disadvantages, such as a tendency for specific kinds of socially acceptable views to surface and for particular members of the group to be overly dominant. (Smithson, 2000, P116)

The student's research project will not require the use of focus groups, as the surveys fulfil the role of collecting the data with respect to Internet users' views on privacy online. Nonetheless, they would prove to be useful in a similar legal research project. In this instance, they could be used to measure the opinions of social media account holders, with respect to what their views are on the growing threat of personal data thief and their online privacy. Furthermore, the focus groups could help the researcher to analyse the account holders' views and opinions on how they feel about their data being used for Big Data purposes.

The following questions could form the basis of the focus group discussion;

Does the fear of personal privacy and data protection online have a negative effect on Internet users and users of social media websites specifically?

Are social media companies and others doing enough to manage the privacy and data protection threats for their account holders or customers?

3.2.3. INTERVIEWS

Students who approach their dissertations by using a socio-legal research methodology will likely use interview or comparative methods, as opposed to quantitative methods.

Bradshaw believes that most students in law are insufficiently accomplished in order to understand and being proactive when dealing with technical issues, such as sampling, experimental design and computer software packages. (Bradshaw, 1997, cited in Thomas, 1997, p103)

An example of the use of an interview research method in law, consisted of a study that concerned a number of family assistance orders, these were made on thirty-five families. As part of the research process, interviews were conducted with six adults, who were named in one of the orders. (Trinder & Stone, 1998) Furthermore, interviews were carried out with nine court welfare officers, who were responsible for each of the orders in the sample. As part of the interview process, officers described and appraised each case and discussed the work relevant to the family assistance orders. (Trinder & Stone, 1998)

The family assistance orders study is an example of how interviews can be very useful in the research process. But, there are limits to be usefulness of the interview method, whether this is by way of a questionnaire or interview specifically. In respect of interviews, it is important to ensure that an appropriate plan is in place, which manages how data was collected. The results of a study by Harris & Brown, show the limitations of the method,

as it was clear that interview data were contextualised and illustrated personal responses provided in a contrived communication setting. (Lankshear & Knobel, 2004).

The student as part of his project has conducted interviews with a number of experts in their fields. These included academics and some business people, who shed light on social media companies plans to deal with privacy issues, for the Internet user. The interview research method played an important role in this research project, as the interviews added real substance to the qualitative data. The purpose of these interviews was to attain the most relevant and up-to-date views and knowledge from experts in their respective fields. Consequently, this information has played a key role in shaping the thesis's outcome.

3.2.4. DISCOURSE ANALYSIS.

There are a number of approaches that researchers can use when analysing their interviewee participants. One such approach is Discourse Analysis, which involves scrutinising the transcribed story for its important elements, and reviewing the words that were spoken by the interviewee. This is achieved by listening to the tone, pitch, pauses and repetitions, as a way to discover the meaning of the text. (Gee, 1991) Discourse analysis can also be used to disentangle conceptual issues, by means of analysing records and reports. (David Silverman, 2011, p95)

The research method, discourse analysis, can be seen in a legal research setting, in the book by David Silverman, 'Qualitative Research, Issues of Theory, Method and Practice'. Silverman illustrates how discourse analysis is used to simplify complex Scottish government health policy documents. (David Silverman, 2011, p96) In what he calls a policy discourse, a wider audience as opposed to just the policy experts, can understand the words contained in these documents. This process is assisted by using data from interviews and political speeches. (David Silverman, 2011, p98)

Obviously, Silverman's example illustrates the benefits of using discourse analysis. But, in other situations the method can prove to be less useful. When using conversation analysis, which is a component of discourse analysis,

An expert in discourse analysis Schiffrin, states that conversation analysis forms its own assumptions and methodology. Consequently, the veracity of the data is questionable, as the data analysis does not incorporate the specific author's views on their interviews and discussions. (Schiffrin, 1994)

The student's research thesis did not require the use of the research method, discourse analysis. This method has proved to be very beneficial for those legal research projects, which consist of very complex legal scenarios, such as projects that are focused on constitutional or tort law. In this instance, the student will be carrying out interviews on the area of data protection law and data protection and privacy issues from both a legal and technology perspective. Also, the level of legal complexity will be relatively low, which will mean that discourse analysis will not be required in this case.

The topic of research has elements of complexity, which could benefit from a discourse analysis approach. However, the other quantitative methods that are being used in the thesis, such as interviews and questionnaires, mean that the project is more transparent and less complex for the perspective reader.

3.3. DETAIL ON THE QUALITATIVE DATA – INTERVIEWS

The questions that were posed to the interviewees, who are experts in IT law and privacy technologies, focused on a number of key issues, which were important in answering the research questions. In the interview section, the author examined the important legal and some technical issues with respect to Big Data.

The most important legal issues concerned consent notices for potential secondary purposes that do not yet exist or have not been conceived; if current anonymisation policies go far enough to protect citizens' personal details; whether consent be attained for forms of data reuse, like data recycling and data sharing and if consent be attained for types of data reuse, such as data repurposing and data recontextualization; whether social media is compatible with privacy; how organisations can demonstrate that consent has been obtained to the standard required by the GDPR, given the likely secondary uses of the data and finally whether valid consent be obtained from data subjects online.

The most important technical or privacy technology issues concerned how the analytics technology can be improved in order to prevent privacy leaks and could attack patterns for de-identification be used?

How do masking integration methods improve differential privacy protection schemes?

3.4. DETAIL ON THE QUANTITATIVE DATA – SURVEYS

The quantitative data results are derived from the three surveys, which questioned Law and IT students and some experts in these fields of research.

The surveys consisted of mainly closed questions, but some of which were open questions. The surveys sought to gauge the views of the students and understand how they feel about their privacy online, especially with respect to online social media. Specifically, for example, the students were asked as to why they use social media and their concerns about not having complete control over the information that they provide online.

3.5. DATA ANALYSIS

Two quantitative analysis software applications were used to assist with the analysis process for the survey results. At the outset, the SPSS software application was used, but this proved to be difficult to use. Because of the small sample size and number of variables. Also, as a result of advice that was received from an academic, whereby it was suggested that the Minitab software application was easier to manipulate and produced less complicated results analysis.

Descriptive statistics were calculated that highlighted the survey respondents' views on their worries with regards to online privacy, their awareness of online privacy and social media settings and their engagement with respect to managing their privacy online. The author used the Chi-Square tests to assess and realise the differences of opinions in relation to privacy online.

The QDA miner software was used to assist in the analysis process for the interviews results.

4. CHAPTER FOUR –RESULTS AND ANALYSIS

The results section of this Chapter consists of both the quantitative and qualitative research. The quantitative results are derived from three surveys, which questioned Law and IT students and some experts in these fields of research. The qualitative results consist of the completed interviews, which include the views and opinions of experts in IT Law and IT.

4.1. MEASURES FOR SURVEYS

The first survey sampled the views and opinions of the IT and Law students at Canterbury Christ Church University, using the Bristol Online Survey software. Forty two anonymous students completed the survey, which consisted of eighteen mainly closed questions, but some of which were open questions. The survey sought to gauge the views of the students and understand how they feel about their privacy online, especially with respect to online social media.

The questionnaire included questions that sought the opinions of the students as to why they use social media and gave the students a choice from, for example; *to connect with friends or I use different social media platforms for different purposes.*

Another question closed question asked the students about their level of concern with respect to not having complete control over the information that they provide online. They were asked; *Would you say you are…?, followed by choices such as; very concerned, fairly concerned and not at all concerned.*

Another open question was designed to seek the express opinions of the forty two respondents with respect *to what worries they have about companies scraping and analysing publicly accessible social media posts?*

The answers varied from one word answers to a number of sentences, most answers provided the author with a valuable and informed insight of their worries about privacy online.

The second survey was completed using the survey monkey software and it consisted of 10 mainly closed questions and some open questions. The aim of this survey was to receive the views and opinions of primarily postgraduate law students and some experts at the

University College Cork in Ireland. Once again, this survey sought to measure the opinions of students on the issue of online privacy and social media privacy. The 26 respondents to the survey monkey questionnaire were asked such questions as, *'Do you believe that your explicit approval should be required before any sort of personal information is collected and processed?'* The multiple choice answers varied from, *'yes, in all situations to yes, when sensitive information is required'*

Another question, which provided the author with evidence of the real opinions of the students, *asked about their concerns with respect to the use of their information by public bodies and private companies.*

The third survey was completed using the survey monkey software and consisted of 10 mainly closed questions and some open questions. The aim of this survey was to receive the views and opinions of students at a University in Sri Lanka, in the area of online privacy in social media. This survey only received responses from the students at the beginning of August, which was during the time of the writing up of the results section of the thesis.

4.2. MEASURES FOR INTERVIEWS

The questions that were posed to the interviewees, who are experts in IT law and privacy technologies, focused on a number of key issues, which were important in answering the research questions. In the interview section, the author examined the important legal and some technical issues with respect to Big Data.

The QDA miner software was used for some of the analysis purposes.

4.3. BOS SURVEY AND TWO SURVEY MONKEY SURVEY RESULTS AND ANALYSIS

RESULTS

The 42 students who completed the Canterbury Christ Church survey were predominately in the less than 20 and the 20-29 age group, and this consisted of 46.7% of the respondents to the survey. The majority of students use social media to connect with friends, at 38.2%

and 29 students. Interestingly, 28.9% per cent stated that they use different social media platforms for different purposes. The questionnaire illustrates that the respondents are aware of the importance of their privacy settings on social media, as the vast majority, 37 students or 88.1% of those surveyed change their privacy settings in order to control who has access to what they post on their respective accounts. Similarly, all of those surveyed are careful when making their personal information accessible on social media.
Half of the respondents state that their contact information is accessible to a select group of people and the remaining 50% are more cautious with respect to their contact information, as they do not post this information on the Internet at all.

There is a change in the cautious approach from the respondents and this can be seen by the manner in which they post items on social media. Because, the majority use their real name when posting items online, that is 29 students or 69%. Although, a minority of this group use a user name for some platforms. Figures 5 to 9 below are the graphic images of the BOS survey user interface and the survey monkey interface.

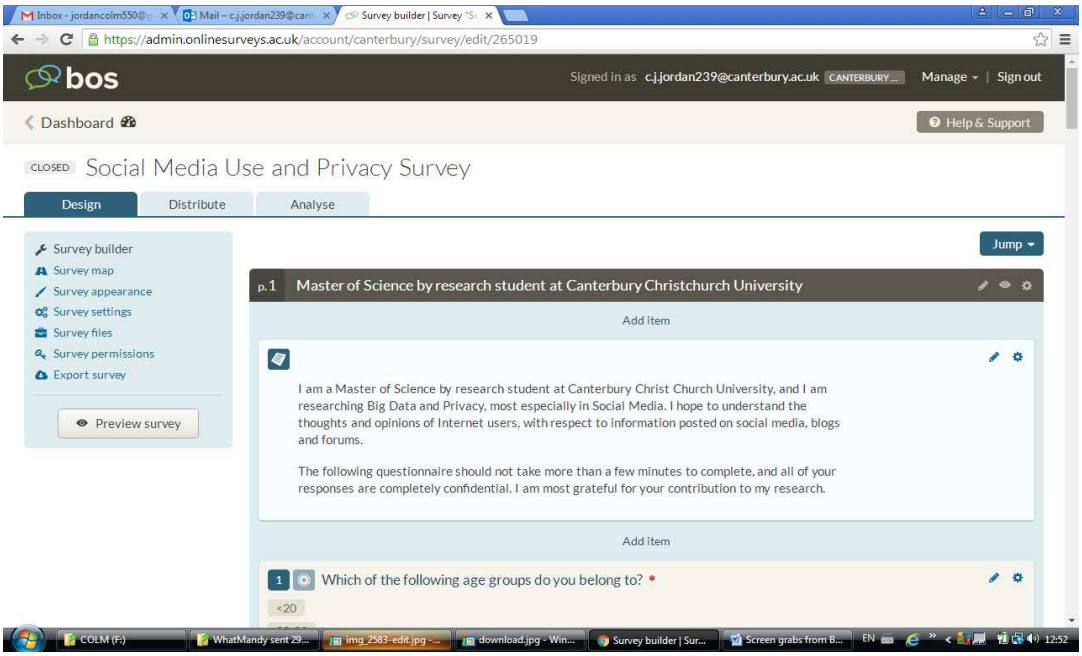


Figure 5: Screen image of the Bristol Online Survey user interface.

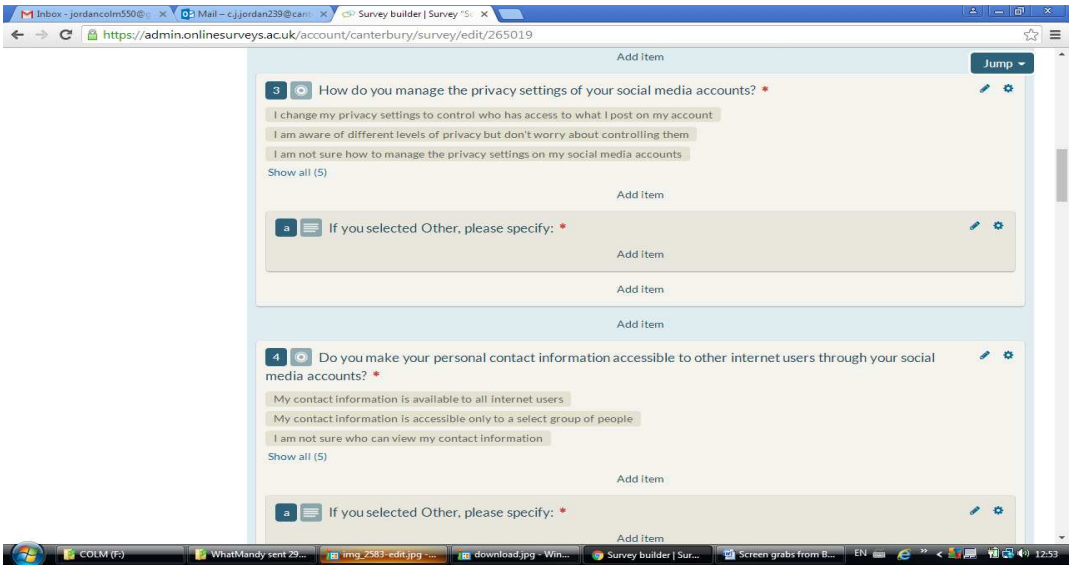


Figure 6: Second screen image of the Bristol Online Survey user interface.

Figure 7: Screen image of the Bristol Online Survey user interface, which includes the editing of the students' questions.

Figure 8: Screen image of the survey monkey online survey user interface, which includes the editing of the students' questions.

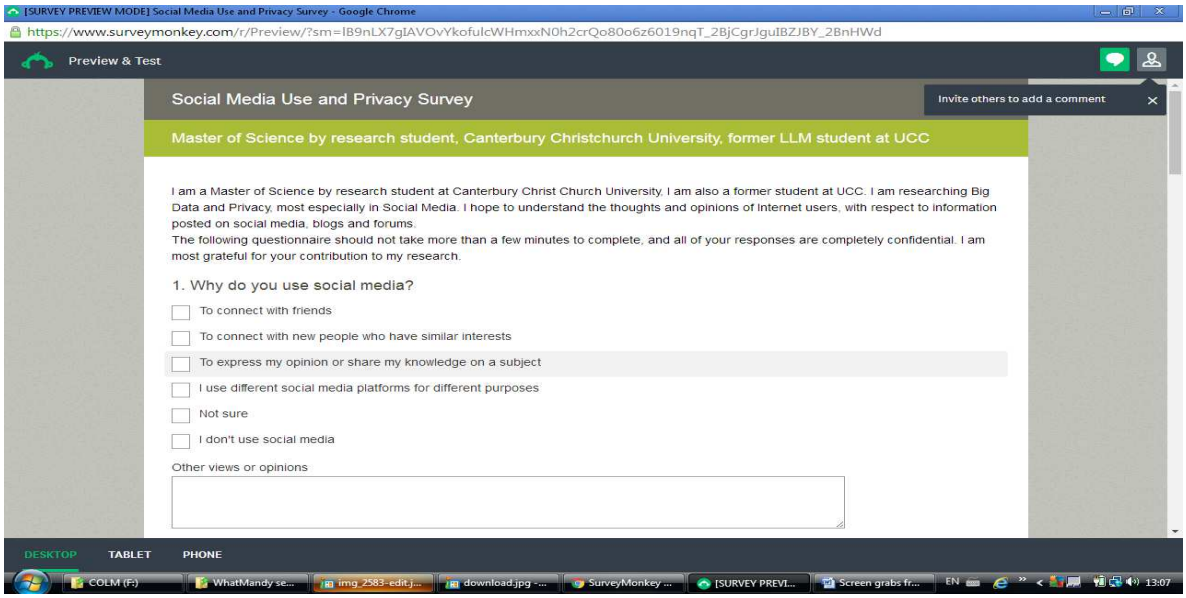


Figure 9: Screen image of the survey monkey online survey respondents' interface.

4.4. SPSS ANALYSIS RESULTS

The following charts represent the data analysis which was carried out using the SPSS software, with the example below, illustrating the comparison of age and privacy concerns. The SPSS software was used for some data analysis purposes, and figures 10 to 13 represent the students' feelings in relation to the use of their personal information by companies. This analysis includes a breakdown of the age groups also, the majority of the students who are concerned about the use of their information belong to the 20 to 29 age group. This includes those who are fairly concerned at 14 students and 11 who are very concerned, that is 25 students out of 42 students in total.

*Do you feel that your explicit approval should be required before any sort of personal information is collected and processed? * Which of the following age groups do you belong to? Cross tabulation*

Count

		Which of the following age groups do you belong to?				
		50-59	30-39	20-29	<20	Total
Do you feel that your explicit approval should be required before any sort of personal information is collected and processed?	No	0	1	1	0	2
	Yes, when sensitive information whether online or offline is required (e.g. health, religion, political beliefs, sexual	0	0	9	1	10
	Yes, when personal information is required online	0	0	7	2	9
	Yes, in all situations	1	3	13	4	21
Total		1	4	30	7	42

Figure 10: Cross tabulation and representation of students' views on collection of their personal information and the corresponding breakdown of age groups.

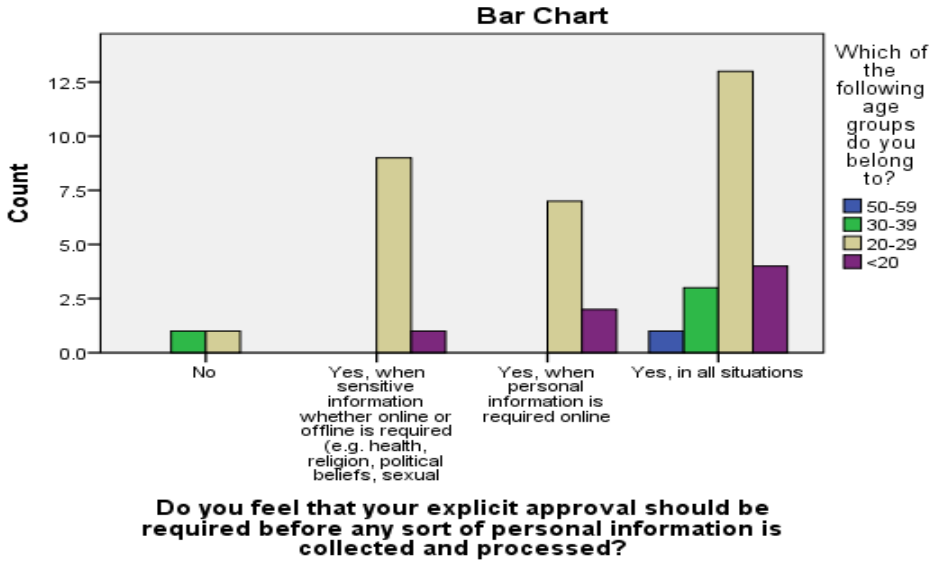


Figure 11: Bar chart illustrating the students' views on the use of their personal information and age classification.

*Public bodies and private companies retaining information about you can sometimes use it for a different purpose than the one it was collected for, without informing you (e.g. for direct marketing, targeted online advertising, profiling). How concerned are * Which of the following age groups do you belong to? Cross tabulation*

Count

		Which of the following age groups do you belong to?				
		50-59	30-39	20-29	<20	Total
Public bodies and private companies retaining information about you can sometimes use it for a different purpose than the one it was collected for, without informing you (e.g. for direct marketing, targeted online advertising, profiling). How concerned are	Not at all concerned	0	0	1	0	1
	Not very concerned	0	0	4	2	6
	Fairly concerned	1	2	14	2	19
	Very concerned	0	2	11	3	16
Total		1	4	30	7	42

Figure 12: Cross tabulation and representation of students' concerns on the use of their personal information and the corresponding breakdown of age groups**.**

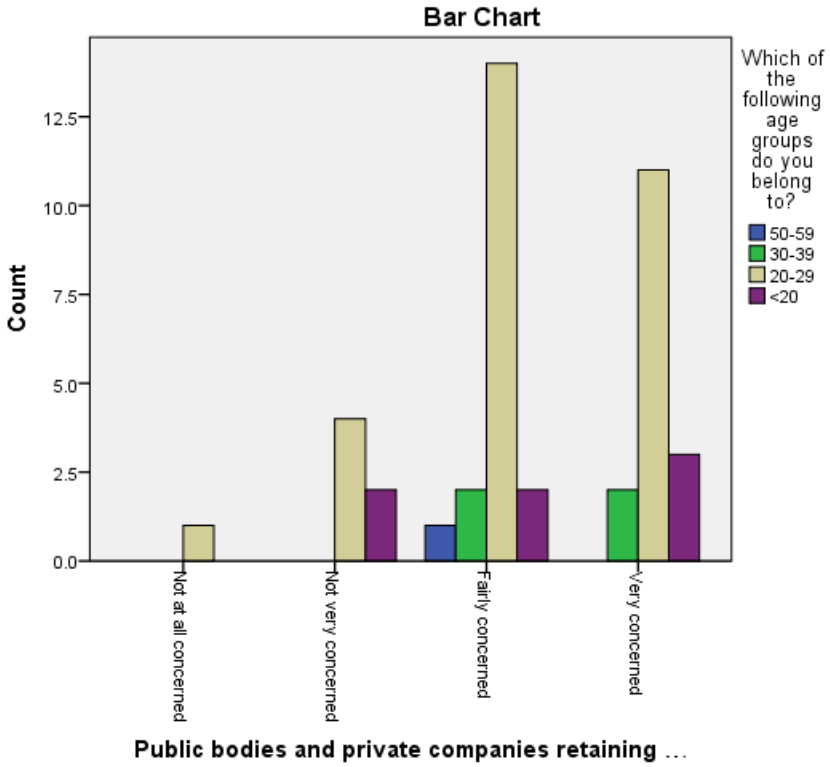


Figure 13: Bar chart illustrating the students' concerns on the use of their personal information and age classification.

4.5. BOS SURVEY RESULTS – ATTITUDES TO SOCIAL MEDIA AND ONLINE PRIVACY

The follow results are the outputs of the analysis of data, performed through the Minitab software application. The first section consists of the students' worries about their privacy online and specifically in social media. It is divided into a number of questions, followed by the results of the analysis. The results include the relevant bar chart of the observed and expected values and the associated Chi Square test results. The Chi Square results will include a discussion of the significance of the p value and others.

The second section includes the respondents' awareness in respect of their online privacy in social media. This section uses the same format as the first section, with a number of questions and the associated breakdown of the results and a discussion of the key values.

The third and final section comprises the survey participants' engagement with privacy issues in relation to their online privacy in social media. Once again, this section uses the same format as the previous two sections, with a number of questions and the associated breakdown of the results and a discussion of the key values.

1. **WORRIES ONLINE WITH REGARDS TO PRIVACY AND SOCIAL MEDIA.**
   a. *Are you worried about companies scraping and analysing your publicly accessible social media posts?*

No –18 Yes – 22. After reviewing the individual answers to the BOS survey question number 12, it revealed that 18 students were not worried about companies scraping their publicly accessible social media posts. But, the majority of 22 respondents were worried by this.

Of those who are worried about privacy, a number of the respondents stated that they believe that it is an invasion of their privacy, they do not like the fact that the big companies are gaining monetary value from their data. Others felt that social media companies are becoming invasive with the practise of scraping their social media posts and that the anonymisation techniques need to be improved.

The respondents who are not worried suggest that the use of the data by the companies will only ensure that the company will provide the social media user with what they want, in respect of the platform. Others believe that their data are being used in any event, as it is in the public domain, but do worry if their private postings may be used.

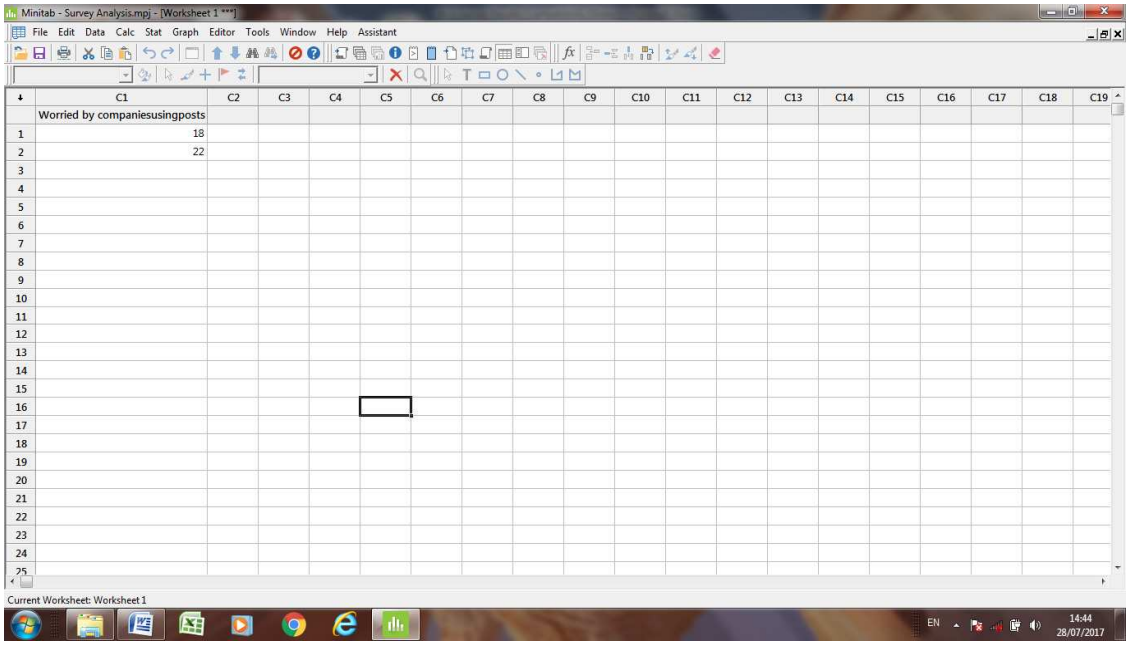


Figure 14: Data entered for minitab analysis.

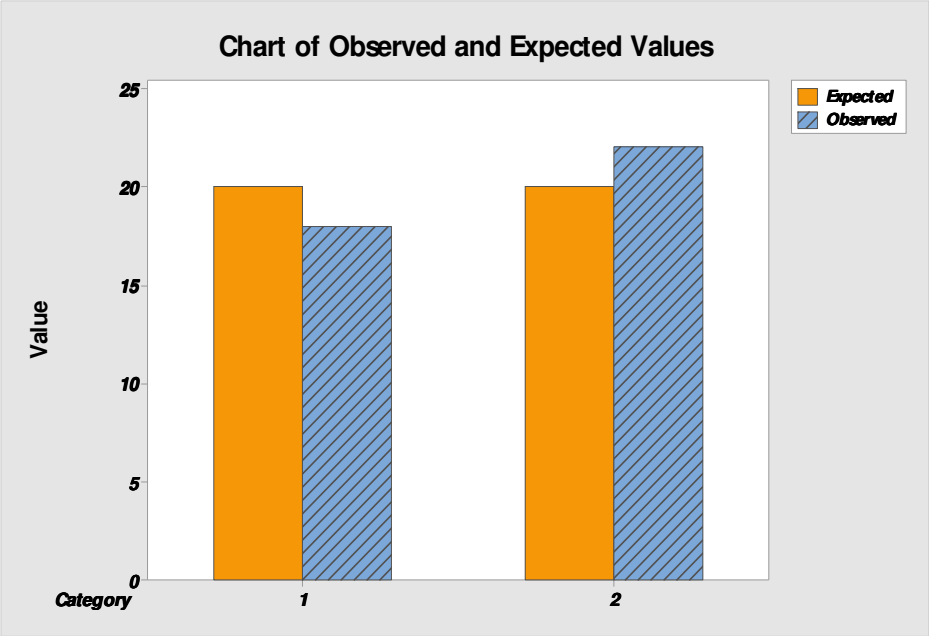


Figure 15: Chart representing the number of those respondents who are worried and not worried about data use online.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	18	0.5	20	0.2
2	22	0.5	20	0.2

Chi-Square Test

N	DF	Chi-Sq	P-Value
40	1	0.4	0.527

Figure 16: The Chi Square goodness of fit test result, for the proportion of those respondents who are worried and not worried about the data use.

42.8% or 18 students are not worried by their data being used, but 52.3% or 22 students are worried by this practise of data reuse. So the Chi Square is 0.4 and the degrees of freedom is 1, as there are 2 groups taking away 1=1, the number of respondents included in the test is 40 or N = 40. The probability is set at < 0.05, but the resulting value is greater than this at < .055.

There is not a significant difference in the number of those who are worried to those who are not, $x^2(1, N=40) =0.4, p < .055$.

1, B. *WHAT MUST COMPANIES SCRAPING FOR ONLINE DATA DO, TO ENSURE SOCIAL MEDIA PRIVACY?*

The answers to the BOS survey question number 13 revealed that 27 students or 38% and the majority of respondents, want companies to make the data anonymous before using it.

While 26 students or 36.6%, feel that companies should ask users' permission to use the data, so the vast majority, 74.6%, want some form of protection measure to be taken before their data is used by the companies.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	11	0.2	14.2	0.7211
2	27	0.2	14.2	11.5380
3	26	0.2	14.2	9.8056
4	6	0.2	14.2	4.7352
5	1	0.2	14.2	12.2704

Chi-Square Test

N	DF	Chi-Sq	P-Value
71	4	39.0704	0.000

Figure 17: The Chi Square goodness of fit test result, representing the breakdown of views on what companies scraping for online data must do to ensure social media privacy.

Firstly, it should be pointed out that question 13, which sought the views on what companies scraping for online data must do to ensure social media privacy, is a more than one answer multiple choice question. Hence, the reason why the N value is 71 as opposed to 42.

The largest contribution to the Chi Square comes from those students who are concerned by the use of their data, a combined figure of 53 students, $x^2(4, N=71) = 39.07$, $p < .001$.

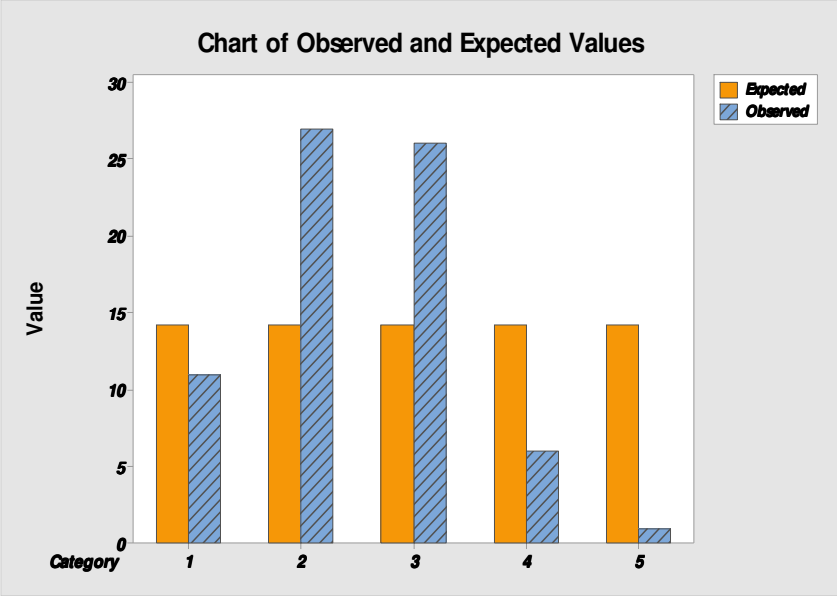


Figure 18: Chart representing the breakdown of views on what companies scraping for online data must do to ensure social media privacy?

2. AWARENESS OF ONLINE PRIVACY AND SOCIAL MEDIA.

a. *How happy are you with the fact that those websites are using information about your online activity to mould or match, advertisements or content to your hobbies and interests?* **Question 14 from the BOS survey.**

Overall the majority of 25 students or 59.6% are not happy with the fact that the websites are using information about their online activity to match web content to their hobbies. But, 13 of the respondents are happy about the information use or 31%.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	2	0.2	8.4	4.87619
2	11	0.2	8.4	0.80476
3	12	0.2	8.4	1.54286
4	13	0.2	8.4	2.51905
5	4	0.2	8.4	2.30476

Chi-Square Test

N	DF	Chi-Sq	P-Value
42	4	12.0476	0.017

Figure 19: The Chi Square goodness of fit test result, representing the breakdown of views on how happy or not the students are with the fact that the websites are using information about their online activity to match web content to their hobbies.

Significantly fewer respondents were very happy for companies to use their information that the other categories, $x^2(4, N=42) = 12.04$, $p < .02$.

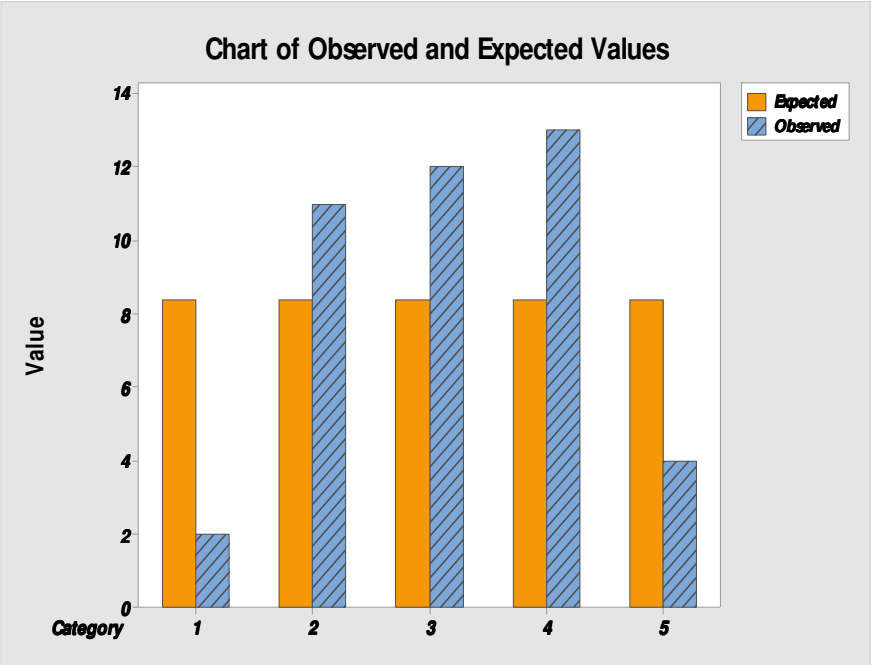


Figure 20: Chart representing the breakdown on how happy the respondents are with the fact that their information is being used by companies.

2.B. BREAKDOWN OF THOSE WHO USE OR DO NOT USE THEIR REAL NAME ONLINE?

No –20 Yes – 19. After reviewing the individual answers to the BOS survey question number 7, it revealed that 20 students do not use their real name online, while 19 do use their real name online.

Of those who do not use their real name online, almost all of the respondents highlighted the importance of remaining private and protecting their identity.

The respondents who do use their real name suggest that it is habit that they have developed, while others want people to know who is posting online and want to look professional.

There is no significant difference between those who use their real name online and those who do not use it, the remaining 3 respondents had no opinion on the issue, $x^2(1, N=39) =.025, p < .88$.

Figure 21: The Chi Square goodness of fit test result, representing the breakdown of the students who use their real name online and those who do not.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	19	0.5	19.5	0.0128205
2	20	0.5	19.5	0.0128205

Chi-Square Test

	N	DF	Chi-Sq	P-Value
	39	1	0.0256410	0.873

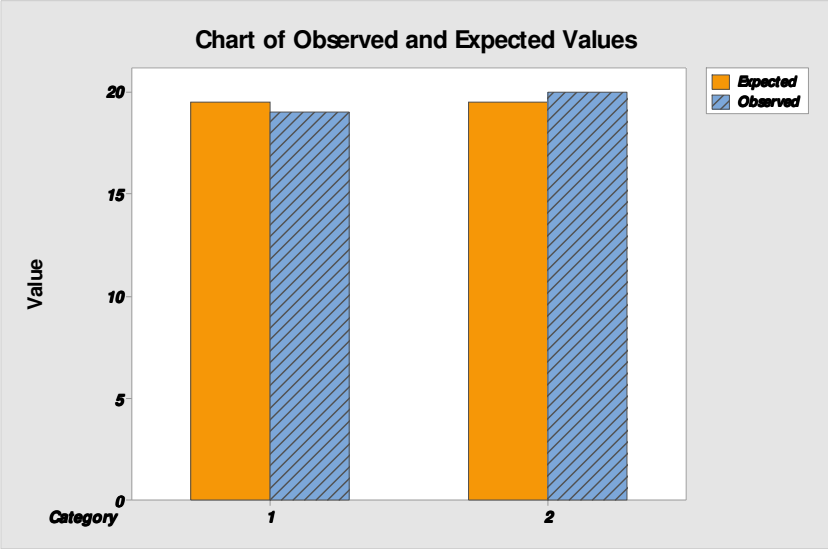


Figure 22: Chart representing the breakdown of the respondents who use their real name online to those who do not.

2.C. BREAKDOWN OF THOSE WHO ARE AWARE AND NOT AWARE AND OTHERS, OF PRIVACY ISSUES AND CONCERNS ONLINE IN SOCIAL MEDIA.

The vast majority of those surveyed, 31 or 73.8% are aware of privacy issues online and most are concerned about protecting their online privacy. While only 3 respondents were not concerned about their privacy online. There was also a fairly high proportion of students who had no opinion on the issue.

Those who were aware and concerned about their privacy online, specifically mention their fears about their contact details being used, others worried about insurance companies using the data to make important decision on claims, and others suggest that people should be provided with some form of education on the issue of privacy online.

Of those who are not concerned, they feel that it is only just data and statistics being compiled, and that no one is specifically being stalked.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	31	0.333333	14	20.6429
2	3	0.333333	14	8.6429

3	8	0.333333	14	2.5714

Chi-Square Test

N	DF	Chi-Sq	P-Value
42	2	31.8571	0.000

Figure 23: The Chi Square goodness of fit test result, representing the breakdown of the respondents who are aware and concerned about their privacy issues online.

There are a significantly higher number of respondents who are aware and concerned about their privacy online and this is having a strong effect on the Chi Square result, $x^2(2, N=42) =31.85, p = 0.00$

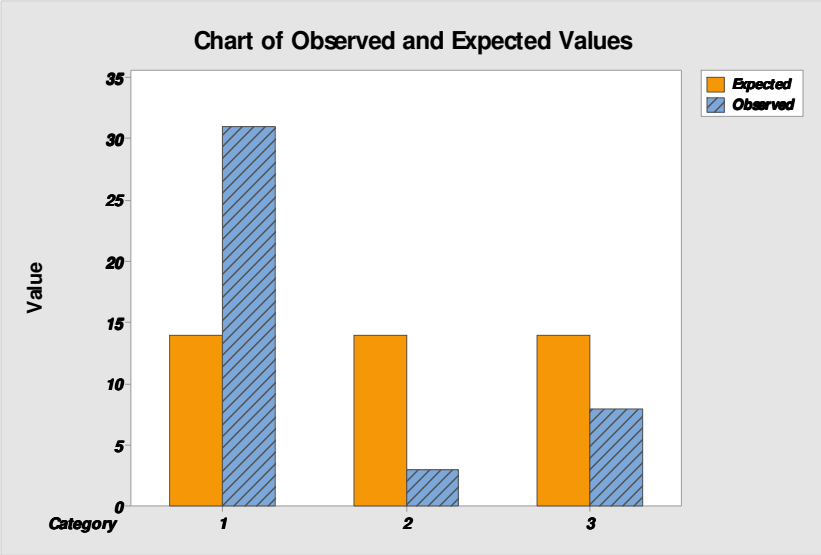


Figure 24: Chart representing the breakdown of the respondents' views on online privacy in social media.

2.D. COMPARISON OF THOSE WHO MAKE THEIR PERSONAL CONTACT INFORMATION ACCESSIBLE TO OTHER INTERNET USERS THROUGH SOCIAL MEDIA ACCOUNTS OR NOT AND THOSE WHO USE OR DO NOT USE THEIR REAL NAME ONLINE.

The was an even split between those respondents' who make their personal data accessible online to those who do not, as it was 21 students each, while as previously seen from the example above, those who use their real online is slightly smaller than those who do not, at 19 to 20.

TABULATED STATISTICS: WORKSHEET ROWS, WORKSHEET COLUMNS

Rows: Worksheet rows Columns: Worksheet columns

	Personal information accessible	Use of real name	All
1	21	19	40
	20.74	19.26	
	0.003241	0.003490	
2	21	20	41
	21.26	19.74	
	0.003162	0.003405	
All	42	39	81

Cell Contents
Count
Expected count
Contribution to Chi-square

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	0.013	1	0.908
Likelihood Ratio	0.013	1	0.908

Figure 25: The Chi Square cross tabulation test result, representing the comparison of the breakdown of the respondents who use and do not use their real name online and those who make their personal data accessible online or do not.

There is no significant difference between those who use their real name online and those who make their personal information accessible online, $x^2(1, N=81) =0.013$, $p = 0.908$.

3. ENGAGEMENT WITH ONLINE SOCIAL MEDIA AND PRIVACY SETTINGS.

A. BREAKDOWN OF WHY THE SOCIAL MEDIA PLATFORM IS USED, USING CROSS TABULATION.

The analysis of why the Social media platform is used, using cross tabulation refers to the more than one answer multiple choice question, number 2. The majority of students use social media to connect with friends at 29 responses or 38.2%. The lowest response rate, which is 6 or 7.9% use social media to express an opinion or to share their knowledge on a subject. Those who do not use social media are not included in the analysis for the question.

TABULATED STATISTICS: WORKSHEET ROWS, WORKSHEET COLUMNS

Rows: Worksheet rows Columns: Worksheet columns

	C1	C2	All
1	29	11	40
	17.25	22.75	
	8.004	6.069	
2	12	28	40
	17.25	22.75	
	1.598	1.212	
3	6	34	40
	17.25	22.75	
	7.337	5.563	
4	22	18	40
	17.25	22.75	
	1.308	0.992	
All	69	91	160

Cell Contents
Count
Expected count
Contribution to Chi-square

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	32.082	3	0.000
Likelihood Ratio	33.982	3	0.000

Figure 26: The Chi Square cross tabulation test result, representing the breakdown of why the Social media platform is used.

The option that had most significance on the Chi Square was to connect with friends, $x^2(3, N=160) = 32.082$, $p = 0.000$.

3.B. CONCERNS ABOUT NOT HAVING COMPLETE CONTROL OVER INFORMATION PROVIDED ONLINE. ALSO, THE OPINION OF THE RESPONDENTS WITH RESPECT TO THE USE OR SCRAPING BY COMPANIES, OF THEIR SOCIAL MEDIA POSTINGS.

As the results of question 10 in the BOS survey illustrate, between those who are fairly concerned, 23 respondents or 54.8% and very concerned, 12 respondents or 28.6%, the majority, 35 students or 83.4%, are concerned about not having complete control over the information they provide online. As the results from question 11 shows, between those respondents who strongly approve, 4 or 9.5% and approve, 12 or 28.6%, the majority, 16 students or 38.1% approve of social media postings being used by companies for marketing purposes. But, a large chuck of the students also disapproves, 12 or 28.6%.

CHI-SQUARE TEST FOR ASSOCIATION: WORKSHEET ROWS, WORKSHEET COLUMNS

Rows: Worksheet rows Columns: Worksheet columns

	Info Control	Postings used	All
1	12	4	16
	8.000	8.000	
2	23	12	35
	17.500	17.500	
3	5	11	16
	8.000	8.000	
4	1	12	13
	6.500	6.500	
5	1	3	4
	2.000	2.000	
All	42	42	84

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	20.015	4	0.000
Likelihood Ratio	22.026	4	0.000

2 cell(s) with expected counts less than 5.

Figure 27: The Chi Square cross tabulation test result, representing the breakdown of the views of students on the control over their data online and the use of information postings online by social media and other companies.

Those who are fairly concerned about not having complete control over their information provided online have had most significance on the Chi Square result, $x^2(4, N=84) = 20.015$, $p = 0.000$.

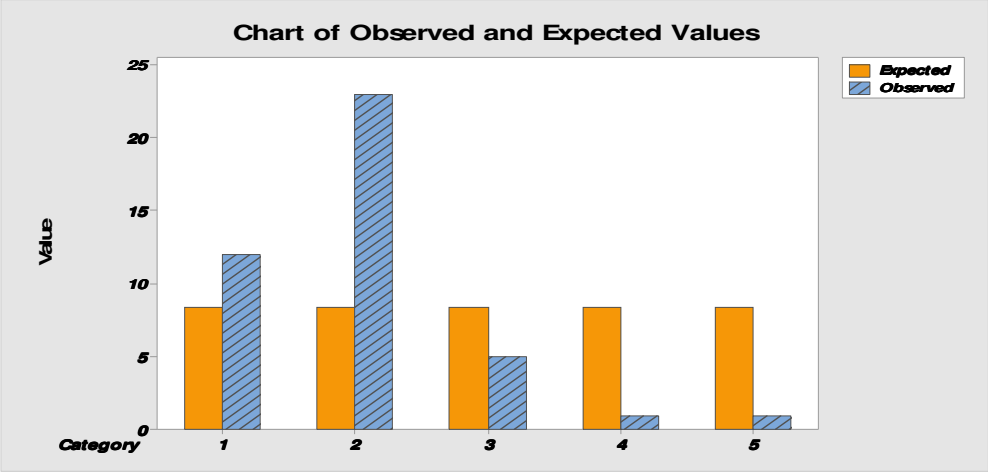


Figure 28: Chart representing the breakdown of the respondents' views on the control over their information online and the use by companies of their postings online.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS IN VARIABLE: C3

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	12	0.1	8.4	1.5429
2	23	0.1	8.4	25.3762
3	5	0.1	8.4	1.3762
4	1	0.1	8.4	6.5190
5	1	0.1	8.4	6.5190
6	4	0.1	8.4	2.3048
7	12	0.1	8.4	1.5429
8	11	0.1	8.4	0.8048
9	12	0.1	8.4	1.5429
10	3	0.1	8.4	3.4714

Chi-Square Test

N	DF	Chi-Sq	P-Value
84	9	51	0.000

Figure 29: The Chi Square goodness to fit test result, representing the breakdown of the views of students on the control over their data online and the use of information postings online by social media and other companies.

Once again, those who are fairly concerned about not having complete control over their information provided online have had most significance on the Chi Square result, $x^2(9, N=84) =51$, $p = 0.000$.

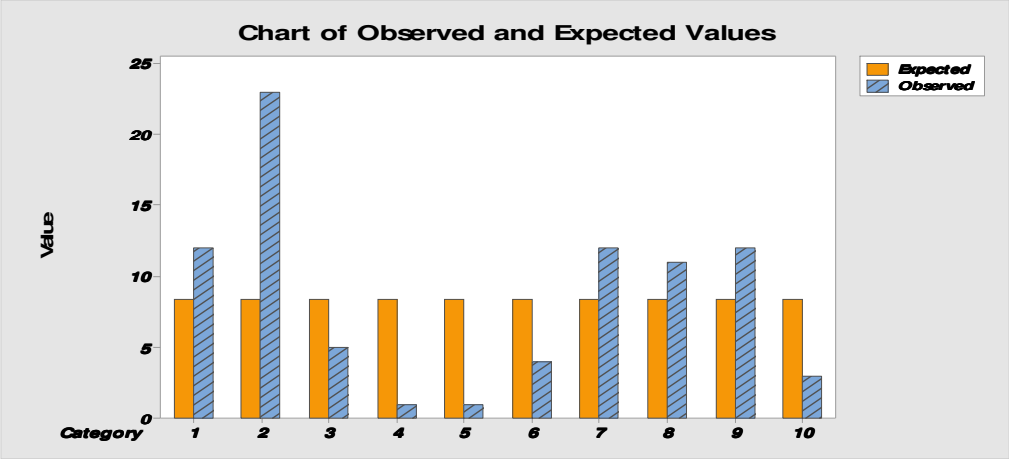


Figure 30: Chart representing the breakdown of the respondents' views on the control over their information online and the use by companies of their postings online.

3. C. DO YOU FEEL THAT YOUR EXPLICIT APPROVAL SHOULD BE REQUIRED BEFORE ANY SORT OF PERSONAL INFORMATION IS COLLECTED AND PROCESSED?

This example is the result of the analysis of question 16 from the BOS survey, and the overwhelming majority at 50% or 21 students feel that their explicit approval is required in all situations, before any form of personal information is collected and processed. Indeed, only 2 students answered no this question or 4.8% of the respondents.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square

1	21	0.2	8.4	18.9000
2	9	0.2	8.4	0.0429
3	10	0.2	8.4	0.3048
4	2	0.2	8.4	4.8762
5	0	0.2	8.4	8.4000

Chi-Square Test

N	DF	Chi-Sq	P-Value
42	4	32.5238	0.000

Figure 31: The Chi Square goodness to fit test result, representing the breakdown of the views of students on whether their explicit approval should be sought before their personal data is processed.

Those students who feel that their explicit approval is required in all situations, before any form of personal information is collected and processed, had the largest effect and significance on the Chi Square result, $x^2(4, N=42) = 32.52$, $p = 0.000$.

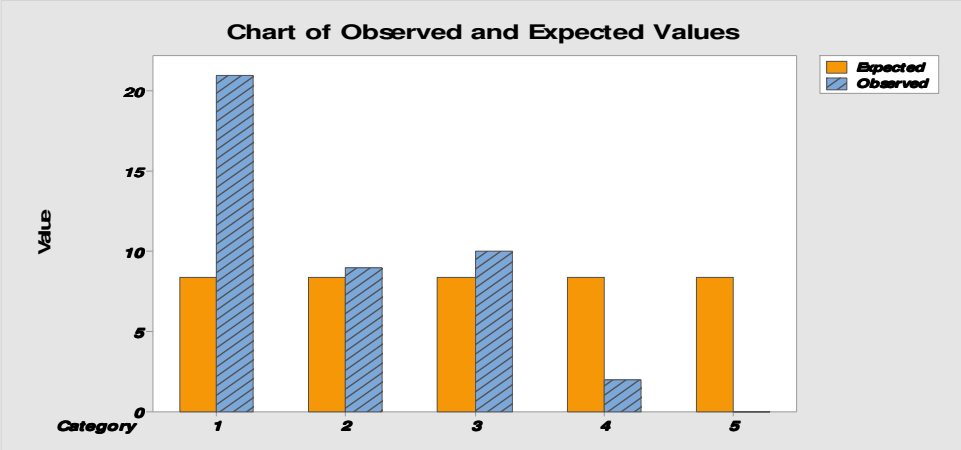


Figure 32: Chart representing the breakdown of the respondents' views on the whether their explicit approval is required in all situations, before any form of personal information is collected and processed.

3. D. INTERNET PRIVACY STATEMENTS DESCRIBE HOW THE PERSONAL INFORMATION YOU SUBMIT WILL BE USED AND WHO WILL HAVE ACCESS TO IT.

IN RELATION TO INTERNET PRIVACY STATEMENTS, WHICH OF THE FOLLOWING BEST DESCRIBES WHAT YOU USUALLY DO?

In relation to Internet privacy statements question number 15, from the BOS survey, the majority of respondents, 21 or 50% of those surveyed, stated that they do not read the privacy statements at all. While, only 4 students or 9.5%, read then in full.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	4	0.25	10.5	4.0238
2	17	0.25	10.5	4.0238
3	21	0.25	10.5	10.5000
4	0	0.25	10.5	10.5000

Chi-Square Test

N	DF	Chi-Sq	P-Value
42	3	29.0476	0.000

Figure 33: The Chi Square goodness to fit test result, representing the breakdown of the views of students, who do read or do not the Internet privacy statements.

Those students who do not read their Internet privacy statements at all had a significant effect on the Chi Square result, $x^2(3, N=42) = 29.04$, $p = 0.000$.

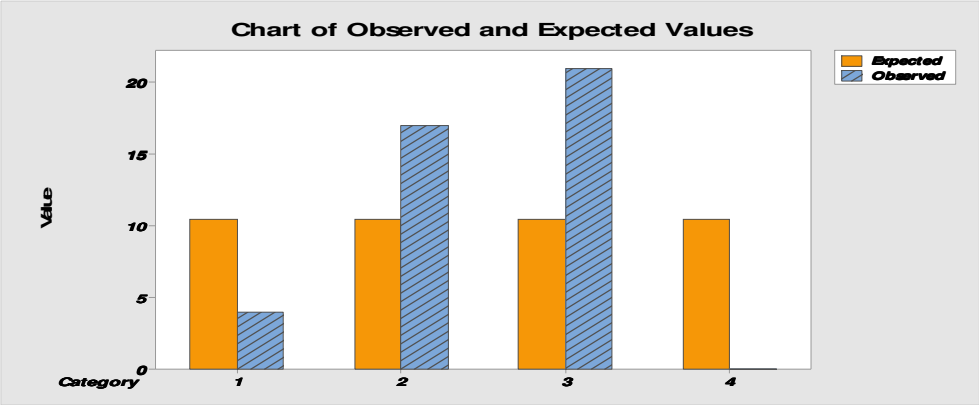


Figure 34: Chart representing the breakdown of the respondents who do read or do not their Internet privacy statements.

4.6. SURVEY MONKEY SECOND SURVEY ANALYSIS

The following is the results analysis of the second survey, which was completed using the survey monkey software and consisted of ten mainly closed questions and some open questions. The aim of this survey was to receive the views and opinions of primarily postgraduate law students and some experts at University College Cork in Ireland, in the area of online privacy in social media**.**

SURVEY MONKEY SURVEY RESULTS – ATTITUDES TO SOCIAL MEDIA AND ONLINE PRIVACY

The follow results are the outputs of the analysis of data, performed through the Minitab software application. This part of the results section is divided into a number of questions. The results include the relevant bar chart of the observed and expected values and the associated Chi Square test results. The Chi Square results will include a discussion of the significance of the p value and others.

EXAMPLE 1 REFERS TO QUESTION 2 IN THE SURVEY, WHICH SOUGHT THE OPINIONS OF THE RESPONDENTS IN RESPECT OF HOW THEY MANAGE THEIR PRIVACY SETTINGS IN THEIR SOCIAL MEDIA ACCOUNTS.

The vast majority of respondents, 18 or 69.23% stated that they change their privacy settings to control who has access to what they post on their account. While the minority of 5 or 19.23% of the 26 students surveyed, stated that they are aware of different levels of privacy but do not worry about controlling them.

Chi-Square Goodness-Of-Fit Test For Observed Counts

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	18	0.25	6.5	20.3462
2	5	0.25	6.5	0.3462
3	1	0.25	6.5	4.6538
4	2	0.25	6.5	3.1154

Chi-Square Test

N	DF	Chi-Sq	P-Value
26	3	28.4615	0.000

Figure 35: The Chi Square goodness to fit test result, representing the breakdown of the views of students, on how they manage their privacy settings in social media.

Those students who change their privacy settings to control the access to their postings online had the most significant effect on the Shi Square result, $x^2(3, N=26) = 28.46$, $p = 0.000$.

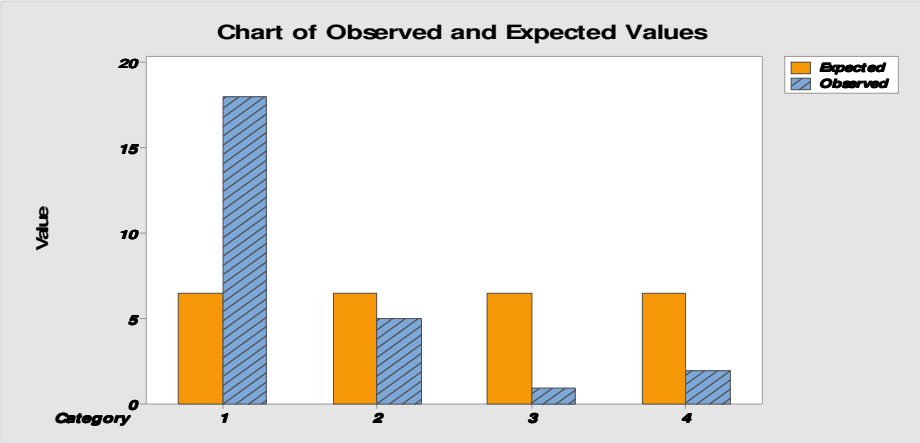


Figure 36: Chart representing the breakdown of the respondents' views on how they manage the privacy settings of their social media accounts.

EXAMPLE 2 REFERS TO QUESTION 3 IN THE SURVEY, WHICH GATHERED THE VIEWS OF THE RESPONDENTS IN RELATION TO THEIR SOCIAL MEDIA MESSAGES BEING VISIBLE TO ALL INTERNET USERS.

The vast majority again of 16 or 61.54% declared that they made some of their social media postings public, but they made information that might be sensitive available only to a select group of people. On the other hand, the minority of 2 or 7.69% made all of their social media postings public, however they did not publish any sensitive information that they did not want other people to see.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	0	0.166667	4.33333	4.3333
2	2	0.166667	4.33333	1.2564
3	16	0.166667	4.33333	31.4103
4	6	0.166667	4.33333	0.6410
5	2	0.166667	4.33333	1.2564
6	0	0.166667	4.33333	4.3333

6 (100.00%) of the expected counts are less than 5.

Chi-Square Test

N	DF	Chi-Sq	P-Value
26	5	43.2308	0.000

Figure 37: The Chi Square goodness to fit test result, representing the breakdown of the views of the respondents, in relation to their social media messages being visible to all Internet users.

Those students who make some of their social media postings public had the most significant effect on the Chi Square result, $x^2(5, N=26) = 43.23$, $p = 0.000$.

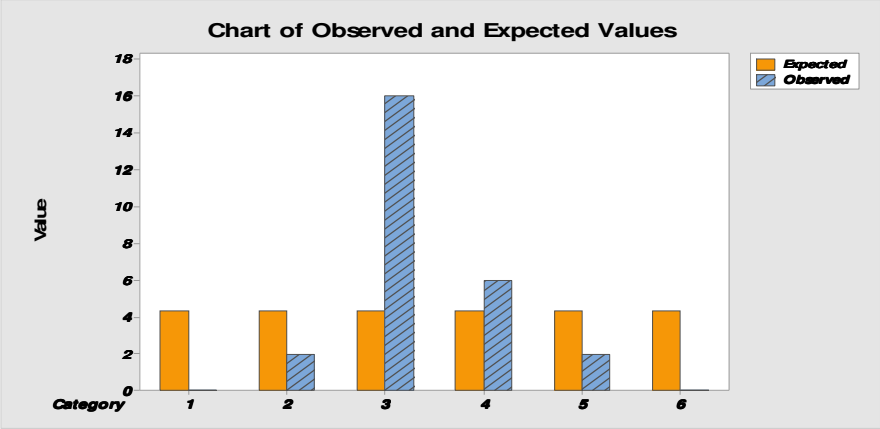


Figure 38: Chart representing the breakdown of the respondents' views, in relation to their social media messages being visible to all Internet users.

EXAMPLE 3 REFERS TO QUESTION 5 IN THE SURVEY, WHICH WANTED TO HEAR THE OPINIONS OF THE RESPONDENTS WITH RESPECT TO THEIR ONLINE POSTINGS BEING USED BY COMPANIES.

There was an even split between those who approved, at 26.9% or 7, of the use of their online postings by companies and did not, at 26.9% or 7 students. But there was a strong response from those who strongly disapproved of the use of their online postings, at 19.223% or 5 respondents.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	1	0.2	5.2	3.39231
2	7	0.2	5.2	0.62308
3	6	0.2	5.2	0.12308
4	7	0.2	5.2	0.62308
5	5	0.2	5.2	0.00769

Chi-Square Test

N	DF	Chi-Sq	P-Value

26 4 4.76923 0.312

Figure 39: The Chi Square goodness to fit test result, representing the breakdown of the views of the respondents, in relation to the use of their social media postings by companies.

Those who strongly approved of their postings being used by companies was significantly fewer than the other response categories and this affected the Chi Square result more than the other categories, $x^2(4, N=26) = 4.76$, $p = 0.312$.

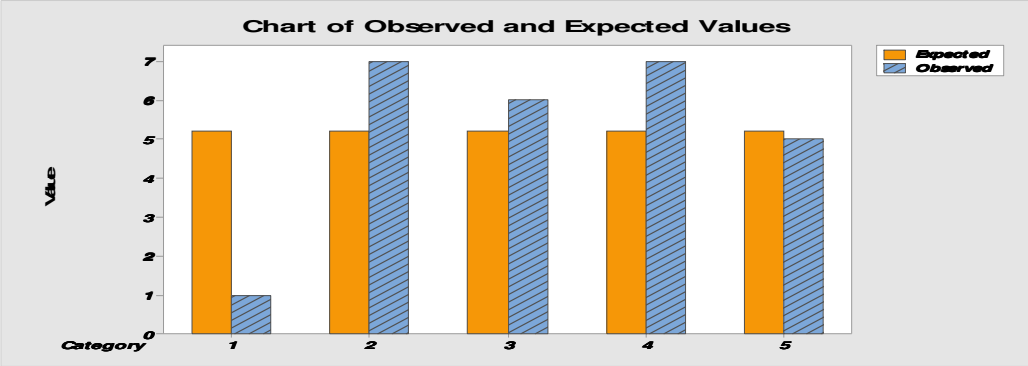


Figure 40: Chart representing the breakdown of the respondents' views, in relation to the use of their social media postings by companies.

EXAMPLE 4 REFERS TO QUESTION 7 IN THE SURVEY, WHICH SOUGHT THE VIEWS OF THE RESPONDENTS ON WHAT COMPANIES SCRAPING FOR ONLINE DATA MUST DO TO ENSURE SOCIAL MEDIA PRIVACY.

The majority of respondents were cautious with regards to their data, as 15 students or 57.69% want the data to be made anonymous before it is used and 17 or 65.3% want their permission to be sought before the data is used.

There was an even split between those who approved, at 26.9% or 7, of the use of their online postings by companies and did not, at 26.9% or 7 students. But there was a strong response from those who strongly disapproved of the use of their online postings, at 19.223% or 5 respondents. While only two respondents were happy for the companies to do nothing if the data was in the public domain. This question was a more than one answer multiple choice question, hence the reason why there are more than 26 responses throughout the response categories.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	2	0.25	8.75	5.20714
2	15	0.25	8.75	4.46429
3	17	0.25	8.75	7.77857
4	1	0.25	8.75	6.86429

Chi-Square Test

N	DF	Chi-Sq	P-Value
35	3	24.3143	0.000

Figure 41: The Chi Square goodness to fit test result, representing the breakdown of the views of the respondents, in relation to what companies scraping for online data must do to ensure social media privacy.

Those who were cautious with respect to what companies must to protect their data had the most significant affect on the Chi Square result, $x^2(3, N=35) =24.31$, $p = 0.000$.

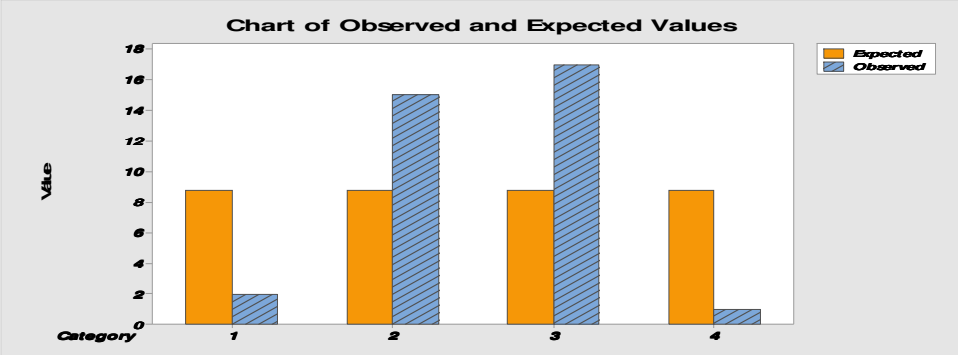


Figure 42: Chart representing the breakdown of the respondents' views, on what companies scraping for online data must do to ensure social media privacy.

EXAMPLE 5. THIS EXAMPLE REFERS TO QUESTION 8 IN THE SURVEY, WHICH SOUGHT TO GAUGE THE LEVEL OF CONCERN THAT THE RESPONDENTS HAVE, WITH REGARDS TO THE FACT THAT PUBLIC BODIES AND PRIVATE COMPANIES USE THEIR DATA FOR DIFFERENT PURPOSES THAN THE ORIGINAL PURPOSE.

Between those who were very concerned, 13 respondents or 50% and those who were fairly concerned, 8 or 30.7%, the vast majority of respondents were concerned with the use of their information for different purposes than the original purpose at the time of collection.

The remaining students, 5 or 19.2% were not very concerned at the use of their data for different purposes.

CHI-SQUARE GOODNESS-OF-FIT TEST FOR OBSERVED COUNTS

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
1	13	0.2	5.2	11.7000
2	8	0.2	5.2	1.5077
3	5	0.2	5.2	0.0077
4	0	0.2	5.2	5.2000
5	0	0.2	5.2	5.2000

Chi-Square Test

N	DF	Chi-Sq	P-Value
26	4	23.6154	0.000

Figure 43: The Chi Square goodness to fit test result, representing the breakdown of the concerns of the respondents, with regards to the fact that public bodies and private companies use their data for different purposes than the original purpose.

Those who were very concerned that they data is used for different purposes had the most significant affect on the Chi Square result, $x^2(4, N=26) = 23.61$, $p = 0.000$.

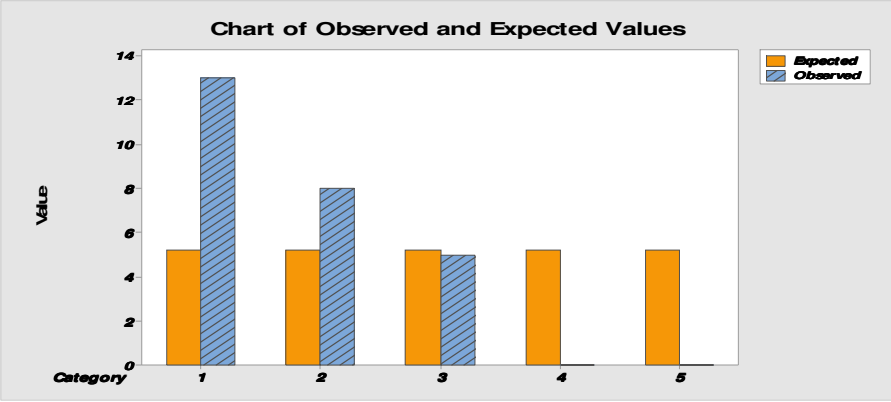


Figure 44: Chart representing the breakdown of the respondents' views, with regards to the fact that public bodies and private companies use their data for different purposes than the original purpose.

4.7. SECOND SURVEY MONKEY SURVEYS RESULTS AND ANALYSIS

The following is the results analysis of the third survey, which was completed using the survey monkey software and consisted of ten mainly closed questions and some open questions. The aim of this survey was to receive the views and opinions of students at a University in Sri Lanka, in the area of online privacy in social media. This survey only

received the responses from the students at the beginning of August, which was during the time of the writing up of the results section of the thesis.

SURVEY MONKEY SURVEY RESULTS – ATTITUDES TO SOCIAL MEDIA AND ONLINE PRIVACY

As of the 5th of August, there were 37 responses to the survey. The follow is a brief discussion of the results of the survey, highlighting the patterns in the survey.

There are many similarities in the opinions of those surveyed in both survey monkey surveys, for some questions the percentage is almost identical in a number of response categories. But, there will be a brief discussion of where there are differences between the responses in some of the questions posed, which are identical in both survey monkey surveys.

EXAMPLE 2 IN THE SECOND SURVEY MONKEY SURVEY ANALYSIS REFERRED TO QUESTION 3 IN THAT SURVEY, WHICH GATHERED THE VIEWS OF THE RESPONDENTS IN RELATION TO THEIR SOCIAL MEDIA MESSAGES BEING VISIBLE TO ALL INTERNET USERS. THE RESULT OF QUESTION 3 IN THE THIRD SURVEY WILL BE COMPARED TO THOSE IN QUESTION 3 OF THE SECOND SURVEY.

In the second survey, the vast majority again of 16 or 61.54% declared that they made some of their social media postings public, but they made information that might be sensitive available only to a select group of people. On the other hand, the minority of 2 or 7.69% made all of their social media postings public, however they did not publish any sensitive information that they did not want other people to see. But, in the third survey, a small majority of 15 out of 37 respondents or 40.5% stated that all of their social media postings are protected and are visible only to a select group of people. Only 13 or 35.1% declared that they made some of their social media postings public, but they made information that might be sensitive available only to a select group of people.

So, it appears that the respondents from the third survey are more aware and cautious with respect to protecting their social media postings online.

EXAMPLE 3 FROM THE SECOND SURVEY, REFERRED TO QUESTION 5 IN THAT SURVEY, WHICH WANTED TO HEAR THE OPINIONS OF THE

RESPONDENTS WITH RESPECT TO THEIR ONLINE POSTINGS BEING USED BY COMPANIES.

There was an even split between those who approved, at 26.9% or 7, of the use of their online postings by companies and did not, at 26.9% or 7 students. But there was a strong response from those who strongly disapproved of the use of their online postings, at 19.223% or 5 respondents.

In the third survey, from the 35 out of 37 respondents, there was a much higher percentage that approved of the use of their online postings, at 15 respondents or 42.8%, as opposed to the 26.9% who approved from the second survey. There was almost double the number who did not care about the use of the postings from the third survey, at 10 respondents or 28.5%, this compares to the 6 or 23% who did not care from the second survey.

So those from the third survey are not as cautious on this occasion about their postings being used as opposed to the respondents from the second survey, who are more cautious.

EXAMPLE 5 FROM SURVEY TWO, REFERS TO QUESTION 8 IN THAT SURVEY, WHICH SOUGHT TO GAUGE THE LEVEL OF CONCERN THAT THE RESPONDENTS HAVE, WITH REGARDS TO THE FACT THAT PUBLIC BODIES AND PRIVATE COMPANIES USE THEIR DATA FOR DIFFERENT PURPOSES THAN THE ORIGINAL PURPOSE.

Between those who were very concerned, 13 respondents or 50% and those who were fairly concerned, 8 or 30.7%, the vast majority of respondents were concerned with the use of their information for different purposes than the original purpose at the time of collection.

The remaining students, 5 or 19.2% were not very concerned at the use of their data for different purposes.

But, the respondents from the third survey, are even more concerned with the use of their information for different purposes than the original purpose at the time of data collection, as 21 of the 37 respondents or 56.7% stated that they were very concerned about this data use, while 14 or 37.8% felt fairly concerned.

So, those from the third survey are clearly more concerned about the use of their data for different purposes at 94.5% as opposed to 80.7% who are concerned from the second survey.

The answers received from both survey respondents to the final question number 10, please express your thoughts and views, which you might have on the topic of online privacy in social media and the commercial use of public social media data, have produced a number of similar comments and thoughts.

Most are concerned about their privacy and the technologies that are in place in order to keep this level of protection at a high standard. But some think it is ok for publically accessible data to be used, as they feel that it will benefit all consumers more generally.

But others suggest that the effects on users' privacy (if personal and sensitive information is used) can be very negative. But also there are positive benefits for the consumers to improve their customer experience. Users' should be informed on how to control their privacy settings and what should be or not posted online.
This respondent believes customers should be able to control what information is used by other third parties, as privacy is a fundamental right.

Others feel that businesses which scrape data online should be held to a very high standard as they should have no automatic right to do so. A distinction may need to be drawn between such scrapers and the social media platform providers themselves. The social media platform provider is providing a service (usually free of charge), and assuming their own terms and Conditions provide (and they capture adequate consents), they should have more liberty to use the data entered on the social media platform provider.

Others expressed a view that there will be occasions where it is lawful and necessary to process personal data without the explicit consent of the data subject, for instance to protect vital interests, for the detection and prosecution of crime, for national security etc.

4.8. COMMENT

FINDINGS AND DISCUSSION

The purpose of the surveys was to gain an insight into the attitudes and opinions of those students and other respondents with respect to social media and privacy.

The questions that were chosen for the BOS and survey monkey surveys, sought to gauge the opinions of the respondents. The questions chosen for the respondents were designed to hear their views on matters that are both topical and important to users of social media generally. For instance, the question on how the users manage their privacy settings in their social media accounts, using their real name when posting things on social media and the question about having control of information that they submit online. These are issues that are important to social media users and also issues that can affect and impact upon their privacy and data protection in an online setting.

The BOS survey software was used to run the survey for the students at Canterbury Christ Church University, this software is more advanced than the free use survey monkey software. For example, the BOS software makes it easier to analyse the survey results, allows for more questions to be asked per respondent and enables open questions to be asked within the main closed question.

The BOS software could not be used to survey the respondents in Ireland or Sri Lanka, it could only be used for Canterbury Christ Church University students. This is why the survey monkey software was used, and two separate SM surveys were posed to the respondents in Ireland and Sri Lanka.

The BOS survey consisted of more questions and longer, more detailed questions. The more powerful BOS software made it possible for this number and detail in the questions. The survey monkey software allows for a maximum of ten questions and does not have the more powerful functionality of the BOS software. There were more open questions included in the BOS survey, which were posed to the students to elicit more information on their feelings about privacy issues.

The students, who participated in the BOS survey at Canterbury Christ Church, comprised of students who were doing IT courses and Law courses. Those students, who were doing IT courses, were biased, as they had covered security and privacy issues in their courses.

The students from Ireland, who completed the survey monkey survey, were not biased, as they were from a legal background. While the students in Sri Lanka, who completed an identical survey monkey survey to their Irish counterparts, were biased in their answers to the survey, as they had completed security and privacy issues in their IT courses at the Sri Lanka University.

The students who had covered security and privacy issues in their courses provided more informed answers to the open survey questions. Their answers were much longer and more beneficial, in some ways, than those who provided less meaningful answers. Nevertheless, those students who provided unbiased answers, gave some insightful answers in relation to their fears and worries about their privacy online. These answers would be representative of most members of the public who use social media.

It was evident from the survey results that the respondents had firm and strong views with respect to their privacy online and in particular with social media. The majority of students were aware of the importance of privacy online and they were keen to manage their privacy settings appropriately. It was clear also, that the majority of students were cautious with respect to their privacy, but there were some anomalies in the results, in that on the one hand most respondents are able to manage their privacy settings and are cautious but in other ways they are not. For example most respondents use their real name when posting messages on social media, this in particular applies to the BOS survey. Using your real name online can be privacy risk, but on the other hand, the social media users may use their real names in order to maximise the usefulness of the networking platform.

As outlined in the analysis of question 2 for the survey monkey (SM) surveys and question 3 for the BOS survey, this question revealed that most students were aware of their privacy settings and were using them.

Questions 4 and 5 from the BOS survey revealed that the respondents were sensible and cautious with their privacy as the majority made their information available to only a select group of people and similarly they made their postings accessible to a select group. This also applies to question 3 from SM survey also, as their postings are for a select group.

Interestingly though, question 6 from the BOS survey, revealed that when respondents post things online in social media, they mainly use their real name, this is 70% or 29 students out of 42, this includes 15 who use their real name and the 14 use their real name for some platforms and a username for others. This illustrates a less cautious approach to their privacy and security online, but the reasons for doing this included wanting to look professional and wanting to find their long lost friends and network to the maximum level.

Furthermore, when asked in the BOS survey, about how much control do you believe you have over the information you submit online, e.g. the ability to correct, change or delete this information? A large portion of the respondents, 47.6% or 20 out of 42, believe that they have partial control or no control at all over their information.

It could be argued that if they didn't provide so much details and information about themselves, this may not be such an important issue.

Question 17 from the BOS survey, again highlights the fact that the respondents are concerned about their data being used for different purposes than originally intended, as between those who were fairly concerned and very concerned about this issue, accounted for 35 out of the 42 people surveyed or 83.3%.

Question 8 from the SM survey was based on the same issue, and the response was similar as in the first SM survey, those who were fairly concerned and very concerned accounted for 80.7% or 21 out of the 26 respondents. But, it was higher for the second SM survey as between the fairly and very concerned respondents; it was 94.6% or 35 out of the 37 respondents.

Furthermore, question 13 from the BOS survey, shows the respondents' awareness of the importance of protecting their privacy online, as they are asked about what must companies scraping for online data do, to ensure social media privacy?

Once again the vast majority both want to make the data anonymous before it is used or chose 'ask the users' permission to use the data' option, these account for 74.6% or 53 respondents. This was a more than one answer multiple choice question.

Question 7 from the SM survey concerned the same issue and the first SM survey follows the same pattern as the BOS survey, as the vast majority again both want to make the data anonymous before using it or ask the users' permission to use the data, this accounts for 31 respondents for the more than one answer multiple choice question. The second SM survey accounted for 36 out of 37 respondents or 97.29%

Finally, the BOS survey question number 16 asked do you feel that your explicit approval should be required before any sort of personal information is collected and processed?

Once again, the respondents' answers showed that they are concerned about this issue, as between those who said yes and those who said yes, if it is collected when personal information or sensitive information is required online, these options both accounted for the vast majority of the 42 respondents surveyed, at 40 students or 95.2%.

The first SM survey follows with a similar pattern of opinion by the respondents from question 9 in that survey, as those who said yes and those who said yes, if it is collected when personal information or sensitive information is required online, accounted for 24 out of the 26 students surveyed or 92.31%. Again the second survey monkey survey followed suit from question 9, as those who stated yes and those who said yes, if it is collected when personal information or sensitive information is required online, this accounted for all 37 students surveyed or 100%.

Interestingly, when the students from the BOS survey were asked if they read the Internet privacy statements, the majority did not. The results from question 15 of the 42

respondents' shows that 21 or 50% do not read then at all and 17 or 40.5% read them partially.

Unfortunately, as the survey monkey surveys only allow 10 questions, the Internet privacy statements question was not included on both SM surveys.

4.9. RESULTS DISCUSSION

It is clear from the results of the surveys that the respondents are concerned about the use of their data and do not approve of its use for other purposes than originally intended. A large portion of the respondents feel that they do not have complete control over their data and this greatly concerns them. But, the fact remains that most of those surveyed use their real name online and post many messages online, perhaps without completely being aware of the consequences.

But, as was noted from the question about the respondents reading Internet privacy statements, most do not do so. This is hardly surprising, as these statements are designed to be long and complex, so that in reality no one reads them.

But, it is time to have a more simplified version of privacy statements and terms and conditions, especially when Internet users freely provide their information and data online. Also, governments and organisations should provide the users with awareness and easy to understand information briefings, in order to reduce these fears and make it a more level playing field for the members of the public using social media.

LIMITATIONS OF SURVEY

While the three surveys comprising a total of 105 respondents, it was hoped to have a higher number, in order to be able to have a greater sample size, to make the measurements more accurate.

It would have been better to have also included the option of choosing a gender question, as this would have provided the author with a further insight, to see for example if males were less concerned about their privacy than females. The survey monkey software only permitted a total of ten questions and other functionality was not available also.

RESULTS AND ANALYSIS OF INTERVIEWS

The author used QDA miner to code the important and key texts from the documented interviews. QDA miner assited the author in retriving and analysing the coded segements of the interviews.

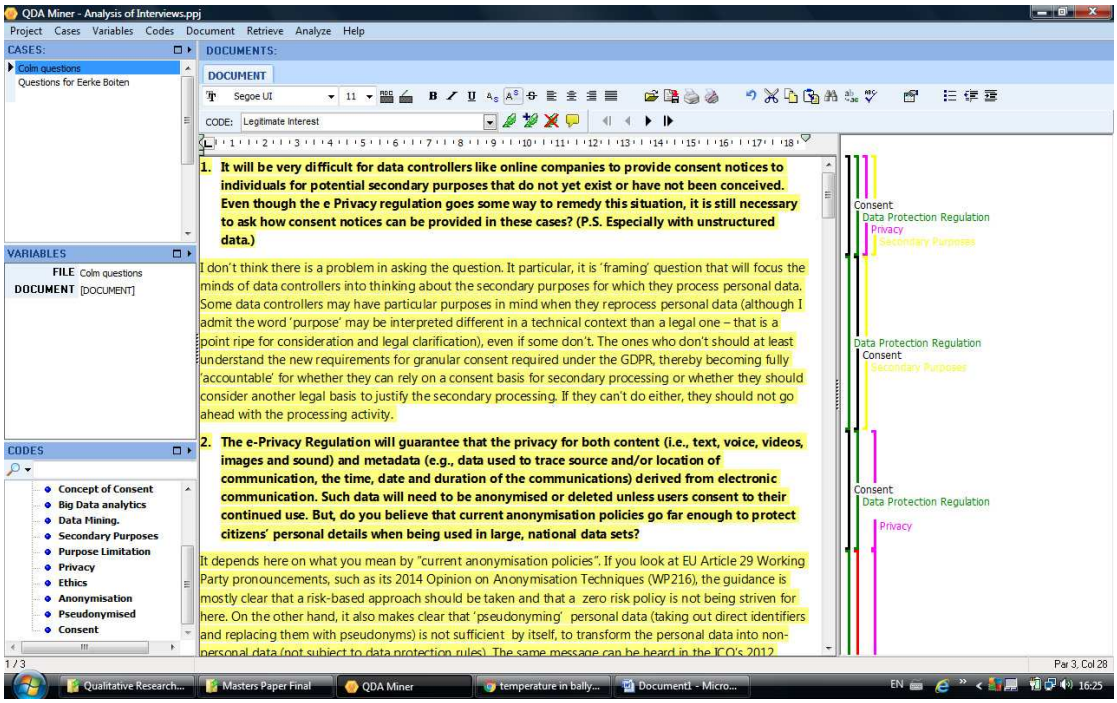


Figure 45: Using QDA miner to code the interview documents.

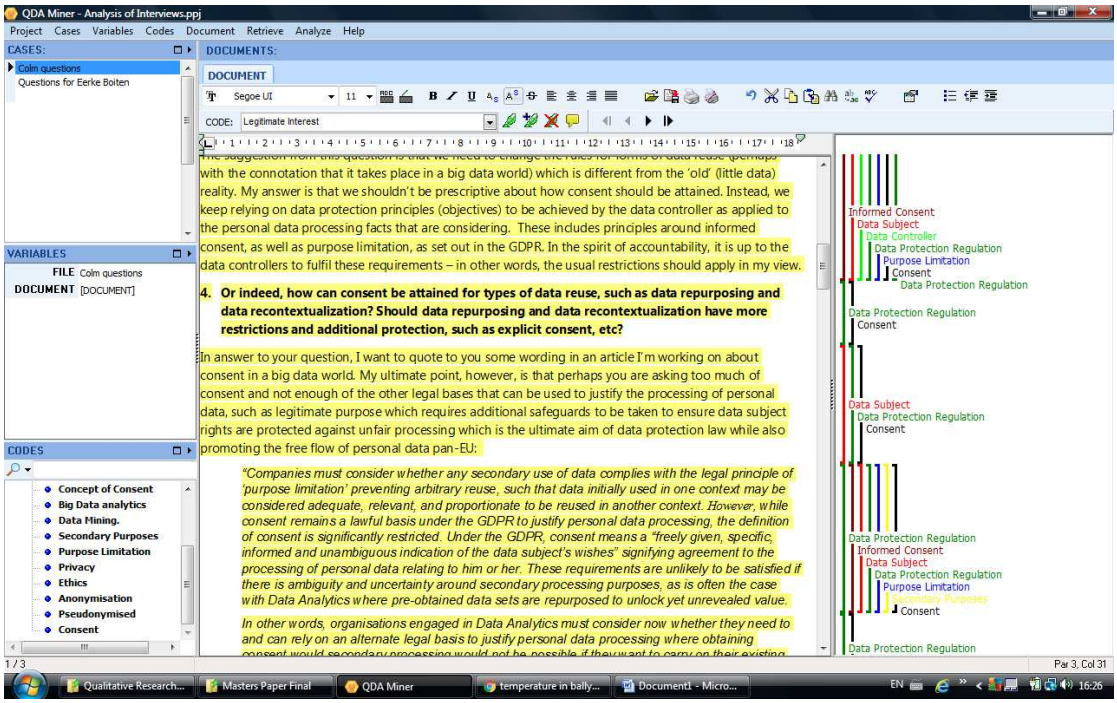


Figure 46: Using QDA miner to code the interview documents.

The bottom left pane in the screen shots from Figure 45 and 46 shows the designated main code and sub codes, which consisted of the important issues from the interviews, such as privacy and consent. The centre pane shows the imported text document, which is the interview document, this is analysed by the QDA miner software. The right pane represents the margin where code marks are located and they indicate the location of the codes in the interview document, for instance Data Protection Regulation is presented in green throughout both figures 45 and 46. The text retrieved function made it possible to produce the codes throughout the interview document. The codes were assigned to the retrieved segments.

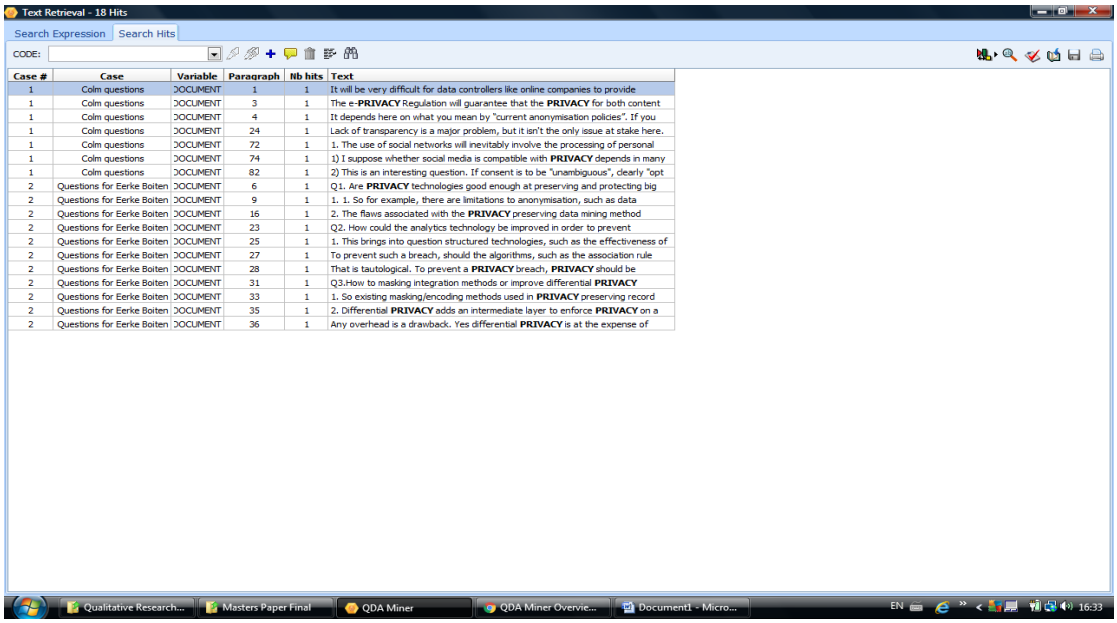


Figure 47: Table of the test retrieval hits for the code or word privacy.

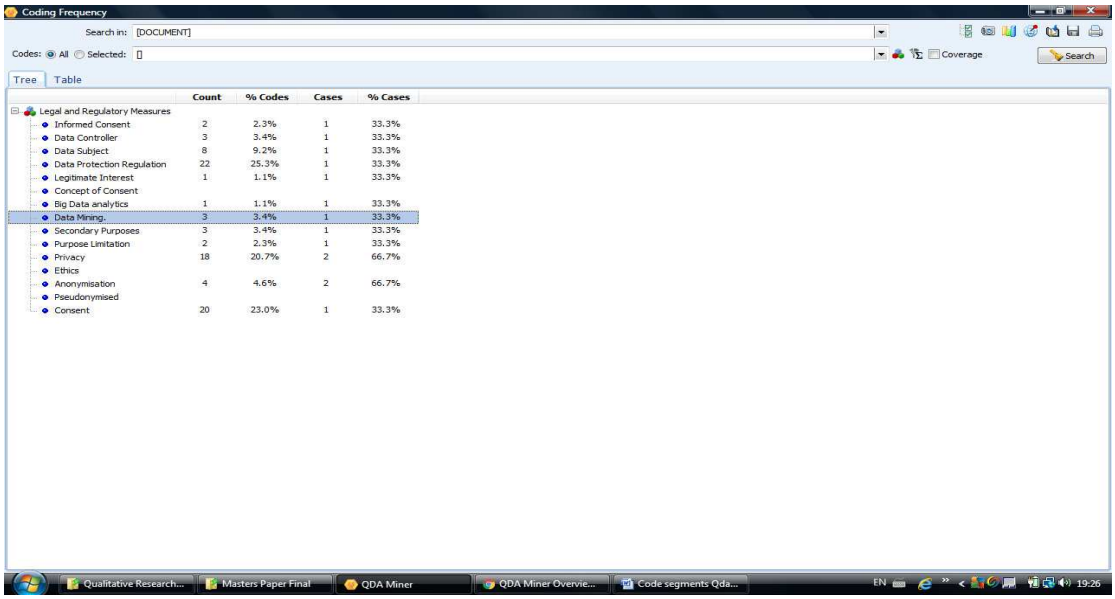
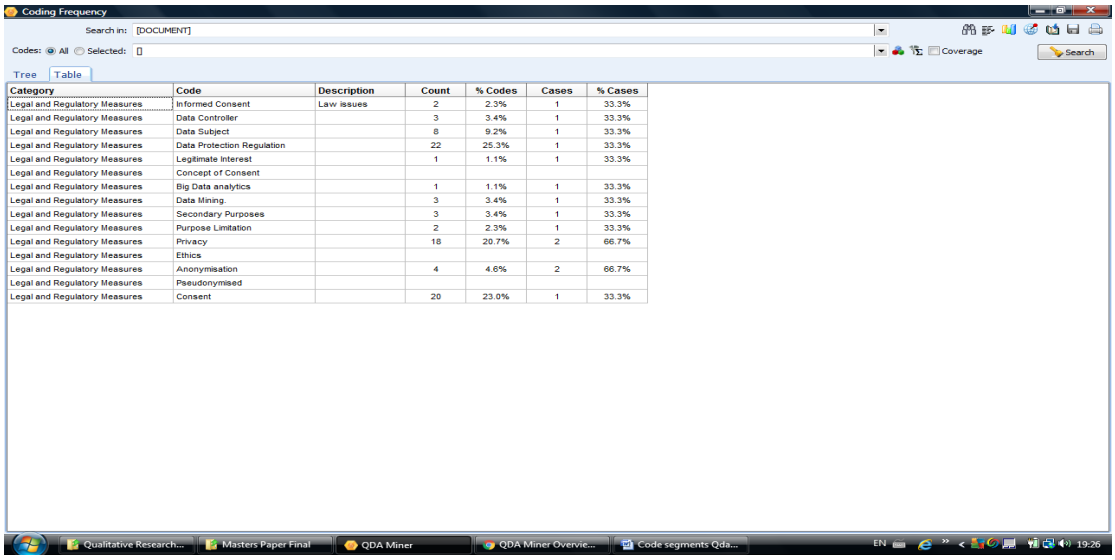


Figure 48: Frequency table representing the quantity of the coded segments, such as privacy and consent.

Using the coding frequency function of QDA miner, the frequency of the code segments can be represented in charts as below. The co occurrence of codes can also be represented by the coding co occurrences function in QDA miner. Multi dimensional scaling plots graphically represent the proximity of codes. The link analysis function enables the visualisation of the connections between codes using a network graph. This is represented in Figure 52 below. It is clear from the results of the analysis that the GDPR and consent are the most common issues discussed by the interviewees, at 25.4% and 23% respectively. These are followed closely by privacy at 20.7%. The main connection between the results from the surveys and results from the interviews is in regards to the importance that consent and privacy means to the students and the experts alike.



Category	Code	Description	Count	% Codes	Cases	% Cases
Legal and Regulatory Measures	Informed Consent	Law issues	2	2.3%	1	33.3%
Legal and Regulatory Measures	Data Controller		3	3.4%	1	33.3%
Legal and Regulatory Measures	Data Subject		8	9.2%	1	33.3%
Legal and Regulatory Measures	Data Protection Regulation		22	25.3%	1	33.3%
Legal and Regulatory Measures	Legitimate Interest		1	1.1%	1	33.3%
Legal and Regulatory Measures	Concept of Consent					
Legal and Regulatory Measures	Big Data analytics		1	1.1%	1	33.3%
Legal and Regulatory Measures	Data Mining.		3	3.4%	1	33.3%
Legal and Regulatory Measures	Secondary Purposes		3	3.4%	1	33.3%
Legal and Regulatory Measures	Purpose Limitation		2	2.3%	1	33.3%
Legal and Regulatory Measures	Privacy		18	20.7%	2	66.7%
Legal and Regulatory Measures	Ethics					
Legal and Regulatory Measures	Anonymisation		4	4.6%	2	66.7%
Legal and Regulatory Measures	Pseudonymised					
Legal and Regulatory Measures	Consent		20	23.0%	1	33.3%

Figure 49: Frequency table representing the breakdown of the quantity of the coded segments, such as privacy and consent.

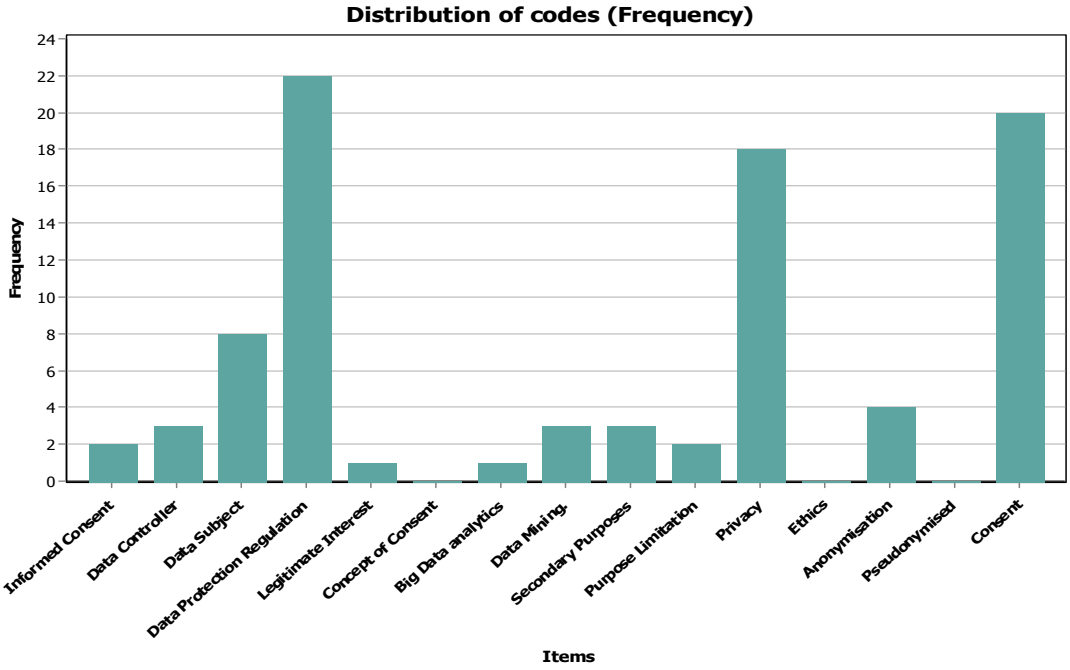


Figure 50: Frequency chart representing the distribution of codes from the interview document.

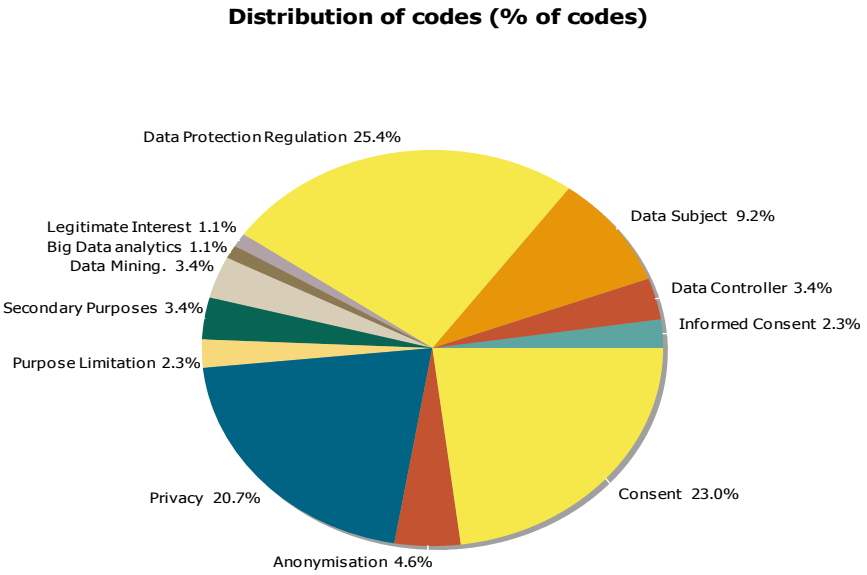


Figure 51: Pie chart representing the distribution of codes from the interview document.

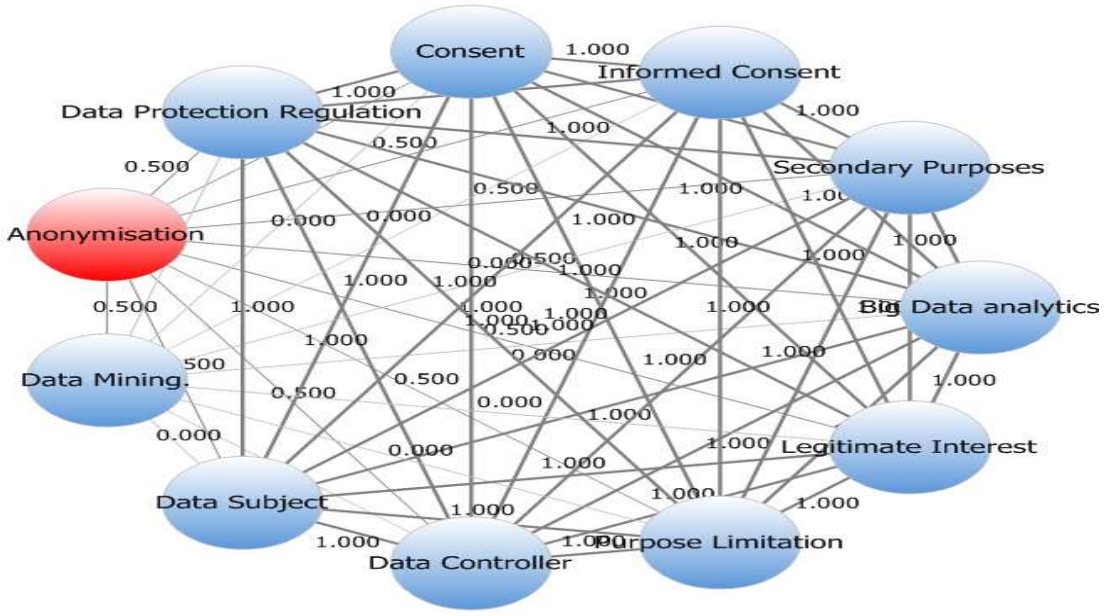


Figure 52: Link analysis function enables the visualisation of the connections between codes using a network graph.

4. 10. QUALITATIVE DATA – INTERVIEWS

METHOD

The questions that were posed to the interviewees, who are experts in IT law and privacy technologies, focused on a number of key issues, which were important in answering the research questions.

The first issue that was put to the interviewees was how data controllers like online companies can, provide consent notices to individuals for potential secondary purposes, which do not yet exist or have not been conceived. While the e Privacy regulation goes some way to remedy this issue, it is still an important and relevant issue.

The interviewees expressed the view that data controllers will have to consider the secondary purposes of the personal data, which they process. Data controllers might have a specific purpose in mind when reprocessing personal data, while the word purpose can be

interpreted differently in a technical context than a legal one. Those data controllers who do not have a purpose in mind should be aware of the new requirements for granular consent that are required under the GDPR. If the data controllers cannot rely on a consent basis for secondary processing or if they cannot consider another legal basis to justify secondary processing, then they should not proceed with the processing activity.

The next issue concerned how to ensure privacy for content and metadata, this type of data needs to be anonymised or deleted unless users consent to their continued use. But, the question was posed to interviewees if they believe that current anonymisation policies go far enough to protect citizens' personal details when being used in large, national data sets.

The interviewees suggested that when looking at the EU Article 29 Working Party pronouncements, such as its 2014 Opinion on Anonymisation Techniques (WP216), it calls for a risk based approach to be taken and not for a zero risk policy. Nevertheless, it clarifies that 'pseudonyming' personal data (taking out direct identifiers and replacing them with pseudonyms) is not sufficient by itself, to transform the personal data into non-personal data, not subject to data protection rules.

The ICO's 2012 Anonymisation Code of Practice and the GDPR also give or make the same declaration. In order to achieve functionally legally anonymised data, including taking account of the exhortations in Recitals 26 of the Directive and the GDPR regarding considering means of identification, requires a context driven analysis of the facts in each particular case. The e-Privacy Regulation is not clear enough about what anonymisation requires in respect of linking it back to the GDPR text. Furthermore, there is a lot of confusion about what terms like pseudonymised and de-identification mean.

The next issue related to consent was also raised, which asked whether consent can be attained for forms of data reuse, like data recycling and data sharing. Also, whether the likes of data recycling, should have little or no restrictions.

This question was answered by the interviewees where they stated that consent should not be attained by a prescriptive format. Alternatively, there should be a reliance on the data protection principles or objectives to be achieved by the data controller, as applied to the personal data processing facts that they are considering.

Such as principles around informed consent, as well as the purpose limitation, as set out in the GDPR. It is up to the data controllers to fulfil these requirements, hence the usual restrictions should apply.

Another issue for the interviewees is related to the previous ones, specifically how can consent be attained for types of data reuse, such as data repurposing and data recontextualization? And whether data repurposing and data recontextualization have more restrictions and additional protection, such as explicit consent, etc?
In answering this, the interviewees suggested that there is too much asked of consent and not enough of the other legal bases that can be used to justify the processing of personal data, such as the legitimate purpose principle, which requires additional safeguards to be taken to ensure that the data subjects' rights are protected against unfair processing, which is the ultimate aim of data protection law, while also promoting the free flow of personal data pan-EU.

Under the GDPR, consent means a "freely given, specific, informed and unambiguous indication of the data subject's wishes" signifying conformity to the processing of data associated to him or her. These requirements are not going to be met if there is ambiguity and uncertainty around secondary processing purposes, as is often the case with data analytics where pre-obtained data sets are repurposed to unlock yet unrevealed value.

Therefore, organisations engaged in data analytics must consider now whether they need to and can rely on an alternate legal basis to justify personal data processing, where obtaining consent for secondary processing would not be possible if they want to carry on their existing business practices going forward.

Essentially, we should not see consent or the other legal bases that might justify secondary processing in isolation. Rather, they should be considered by data controllers upfront and 'in the round' with other principles by which accountability, (the GDPR's focal principle) can be demonstrated. These include data protection impact assessments, data protection by design and default, data transparency (i.e. giving clear and adequate information to data subjects about processing activities planned whatever legal bases is relied on unless exceptions apply), and record keeping (as evidence of compliance risk assessment and mitigation being carried out).

The next issue concerned the lack of openness and information on how data subjects' data is compiled and used, which may mean that they fall victim to decisions that they do not comprehend and have no control over. With this is mind, the experts opinion was sought regarding the lack of transparency for the data subject, with respect to their data?

The interviewees expressed the view that the lack of transparency is a major problem, but it isn't the only issue at stake here. The imbalance in power between data subjects, for example citizens, and major corporations and states is, it could be argued, a bigger question. For example it is a positive step that under pressure from governments, civil society groups and the media, some of the major Internet companies have made their terms of service easier to understand and prompt users to regularly review their privacy settings etc. But this minor concession by the companies doesn't wipe out the imbalance in access to resources and knowledge in the case of a dispute between an individual user and the company.

A further issue for the interviewees concerned how Big Data processes involve working with aggregated data, general patterns and group profiles and to a lesser extent involve the processing of data at an individual level. With this in mind, the interviewees were asked whether they believe that the regulatory framework should continue to centre on the individual and on personal data.

The interviewees answered no, and that it should focus on general interests and rights of groups or class actions. Data are not collected in relation to a specific individual or group (i.e., someone who may have carried out a criminal offence). Instead, they are collected about an indeterminate amount of persons throughout an indeterminate phase of time with no pre-determined reason.

The data are processed on an aggregated level and the profiles revolve around groups, as opposed to certain persons. Persons are judged as a result of pre-determined profiles and pre-determined personality qualities, while the harm to a certain individual is hard to demonstrate.

So, there should really be a group privacy legal regime, a narrow focus on the legal domain is inadequate to deal with most of the difficult issues with human rights violations in the Big Data era.

In addition, the interviewees were asked in the age of Big Data, how can individuals demonstrate personal injury or specific interest in a case, as individuals are frequently oblivious that their rights are being infringed?

Demonstrating that individual users have suffered harm can be difficult, but it does not relieve companies or research institutions from their ethical duties to try and avoid doing harm. The Nuffield Council on Bioethics has a good set of principles on data-driven research in biomedicine, which is a good example of how research can be carried out ethically.

In relation to consent, the question was asked about how can organisations demonstrate that consent has been obtained to the standard required by the GDPR, given the likely secondary uses of the data? (Especially in light of Unstructured Social Media Data)

(As organisations are required to take considerable steps, so as to make sure that data subjects are correctly informed of the purposes for the use of their data).

The interviewees expressed the view that consent is an outdated and obsolete notion. It plays a minor role in the GDPR. The same counts for individual rights. There should be more of an emphasis on the other legal principles like the legitimate purpose principle.

Thinking about it more deeply, if consent is to be "unambiguous", clearly "opt out" modes of consent will not be valid going forward. To date tick box or "I agree" modes of notice and consent have been considered enough to satisfy data protection law's consent requirement, and despite numerous reservations being expressed about the adequacy of this arrangement, it seems likely that it will continue to suffice so far as data protection compliance is concerned.

So far as secondary uses of data is concerned, as many of these secondary uses will not be predictable at the time data is collected from users, it seems impossible that service providers will be able to provide notices or terms and conditions, which explain to their users how their personal data will be used, which in theory will make genuine unambiguous consent impossible also.

Alternatively, very general and wide-ranging privacy notices could be used, but this would be no good either, as consent can hardly be unambiguous if it is given to an unknown. Technological feasibility (or the lack thereof) notwithstanding, the best type of solution would be the inclusion of facilities which allow individuals to track their personal data after sharing with a third party, meaning that if there comes a time that the data are used for a purpose with which the individual does not agree, they can withdraw consent and exercise their right to erasure. This may be pie in the sky, however, not just because of issues relating to whether this is technologically possible or realistic, but also due to the fact that many individuals will simply not have the time, interest or expertise to utilise such facilities.

The next question for the interviewees asked for their view on how valid consent be obtained from data subjects online.

(Particularly in circumstances where there is the complexity of the analytics process, and individual's unwillingness to read terms and conditions. Also, as it may be impossible to determine at the beginning all the purposes for which the data will be used).

The view of one expert was to state No and that he did not even understand what is happening on the Internet with his own data, and as he stated, he is supposed to be an expert.

The alternative view of another expert was to suggest that in order to achieve anything like a satisfactory level of consent, it can be achieved through the incorporation of technological tools like personal data stores, blockchain, and other transparency enhancers, which give individuals the option of visualising their data use, and being able to follow their data after they change hands.

This is far from a panacea to any problems associated with consent for instance, as at present people do not read terms and conditions. So there is no guarantee that they would suddenly decide to use any technological empowerment tools they are given, not to mention the limitations of such technologies. But, it would at least be something, and would hopefully represent an improvement on the models of obtaining consent that are widely used at the moment, which are mostly useless.

What might be worth pursuing is a mode of data protection law that focuses more on data uses rather than data collections. For instance, regulators should target data processing operations, such as automated profiling or Big Data analytics etc., and regulate these more stringently than choosing to focus on individual empowerment etc.

Given the complexity of these data processing operations, we can legitimately ask why the individual should play such an important role in the regulatory framework when they themselves will likely be unable to comprehend what they are being asked to do. In a similar vein, the consequences of these data processing activities are likely to have consequences for groups and society at large as much as individuals, so we might also question whether, given what is at stake, whether individual consent is becoming a proverbial red herring.

The next question sought the views of the experts on how should the individual's right to give their point of view and contest a decision as regards profiling be given effect by a data controller.

In a short answer, one expert stated that they should not be given the right to contest a decision. Instead it should be left to data protection authorities and civil society organisations to enforce.

The use of social networks will unavoidably result in the processing of personal data and therefore employ privacy and data protection laws. Is social media compatible with privacy?

The interviewees stated that the issue of whether social media is compatible with privacy depends in many ways on which conception of privacy one chooses to adhere to. If we consider privacy to be a right to be let alone, the answer is perhaps no, as social media usage necessarily entails some sort of exposure to others, either to other individuals or to service providers like Facebook and Google etc. If, however, you define privacy as the ability for one to control one's information, there might be more scope for compatibility. There may well be other conceptions that can be considered here which may also lead to different answers. But regardless, social media evidently poses considerable challenges to privacy, irrespective of whether privacy is truly incompatible with it.

PRIVACY TECHNOLOGY QUESTIONS.

Q1. Are privacy technologies good enough at preserving and protecting Big Data privacy and preventing information leakage?

1. So for example, there are limitations to anonymisation, such as data subjects lack of control over their data, adversary's background and data transparency to users, i.e.; how much information can be provided to the user on the masking methods and factors used to anonymise a data discharge?

Also, the K-anonymity model cannot offer protection from attribute disclosure via a background knowledge attack or homogeneity attacks.

2. The flaws associated with the privacy preserving data mining method algorithms, such as performance issues in respect of the current privacy preserving data mining methods that show that they are not very effective.

Some of these issues are the effectiveness of data, scalability, the reliability of the data mining and overhead performance.

3. The Limitations of the Hadoop system Software, such as too much duplication of Big Data, data execution that wastes resources and the requirement for specific skills to use Hadoop.

In answering this question, the expert stated that for privacy technologies that are in existence, it is difficult to say in general if they are good enough at protecting Big Data privacy, but it is likely that they are not good enough. For technologies that are in actual use, they are definitely not good enough.
The data subjects' lack of control over their data has nothing to do with anonymisation, but about privacy as control versus privacy as confidentiality. The level of transparency is low in practise, but more is required under the GDPR, especially as pseudonymised data will remain as personal data.

The K-anonymity model looks at a single database on isolation, without considering external resources that might de-anonymise.

Q2. How could the analytics technology be improved in order to prevent privacy leaks?

Could attack patterns for de-identification be used?
So that the linking attack cannot occur, which enables sensitive data to be connected to an individual by combining data from multiple sources, including external sources that are available to the public.

In answering this question the interviewee stated that the attack patterns for de-identification could help create an intermediate between fully supervised access and copying of databases.

Q3.How do masking integration methods improve differential privacy protection schemes?

The expert expressed the view that differential privacy adds an intermediate layer to enforce privacy on a data set, which can result in an overhead that could be considered a drawback, with regards to an in-memory based real-time scenario, in which every single millisecond counts. The most important aspect is that differential privacy can actually change the query results, which is not always the desired outcome. The overhead is a drawback and differential privacy is at the expense of accuracy.

4.11. FINDINGS AND DISCUSSION

In the interview section, the author examined the important legal and some technical issues with respect to Big Data.

The most important legal issues concerned consent notices for potential secondary purposes that do not yet exist or have not been conceived; if current anonymisation policies go far enough to protect citizens' personal details; whether consent be attained for forms of data reuse, like data recycling and data sharing and if consent be attained for types of data reuse, such as data repurposing and data recontextualization; whether social media is compatible with privacy; how organisations can demonstrate that consent has been

obtained to the standard required by the GDPR, given the likely secondary uses of the data and finally whether valid consent be obtained from data subjects online.

The experts' opinions were very helpful and valuable, they also provided an insight into the reality that Big Data has given rise to a plethora of complexity when resolving citizens' personal data rights and issues. Specifically, it is clear that data controllers need to consider a consent basis for secondary processing or another legal basis for doing so, under the the new requirements for granular consent required under the GDPR.

The Information Commissioner's Office (ICO) 2012 Anonymisation Code of Practice and the GDPR are portraying the same message, in that achieving a functional legally anonymised data requires a context-driven analysis of the facts in each case.

There should be a continued reliance on data protection principles or objectives to be achieved by the data controller as applied to the personal data processing facts that they are considering. These include principles around informed consent, as well as the purpose limitation principle, as set out in the GDPR.

Consent or the other legal bases that might justify secondary processing should not be seen in isolation. But, they should be considered by data controllers upfront and together with other principles by which accountability (the GDPR's focal principle) can be demonstrated. These include data protection impact assessments, data protection by design and default, data transparency (i.e. giving clear and adequate information to data subjects about processing activities planned whatever legal bases is relied on unless exceptions apply), and record keeping (as evidence of compliance risk assessment and mitigation being carried out).

Social media evidently poses considerable challenges to privacy, irrespective of whether privacy is truly incompatible with it.

In relation to how organisations can demonstrate that consent has been obtained to the standard required by the GDPR, given the likely secondary uses of the data, a novel suggestion was made that the solution could be the inclusion of facilities which allow

individuals to track their personal data after sharing with a third party, meaning that if there comes a time that the data are used for a purpose with which the individual does not agree, they can withdraw consent and exercise their right to erasure.

In respect of how valid consent be obtained from data subjects online, it was suggested that the best way to achieve anything like a satisfactory level of consent is through the incorporation of technological tools like personal data stores, blockchain, and other transparency enhancers, which give individuals the option of visualising their data use, and being able to follow their data after they change hands.

Finally, the main connection between the results from the surveys and results from the interviews is in regards to the importance that consent and privacy means to the students and the experts alike.

PRIVACY TECHNOLOGY RESULTS

The most important technical or privacy technology issues concerned how can the analytics technologies be improved in order to prevent privacy leaks and could attack patterns for de-identification be used?

In answering this question the interviewee stated that the attack patterns for de-identification could help create an intermediate between fully supervised access and copying of databases.

How do masking integration methods improve differential privacy protection schemes?

The expert expressed the view that differential privacy adds an intermediate layer to enforce privacy on a data set, which can result in an overhead that could be considered a drawback with regard to an in-memory based real-time scenario in which every single millisecond counts. The most important aspect is that differential privacy can actually change the query results, which is not always the desired outcome. The overhead is a drawback and differential privacy is at the expense of accuracy.

LIMITATIONS

The main limitations in the interview and qualitative section of the research were the number of interviewees and the responses that were received from some of those who agreed to participate.

Attempts were made to seek interview participants and contacts were gathered, at the beginning of the research process. Some of those who agreed to be interviewed failed to partake in the process, which was disappointing.

Others, who were contacted and asked if they could help in any way with the research, ignored these requests completely. Then some of the interview participants provided the author with very brief or unhelpful answers.

So, these experiences in the data collection phase of the research project, have been an eye opener and have provided the author with a realisation of what the research process is really like and the potential pitfalls that can occur along the journey.

5. CHAPTER FIVE - CONCLUSION

The aim of the thesis was to examine the Big Data analytics technologies and the associated privacy protection methods, and to determine if these technologies are fit for purpose. While simultaneously, examining the current legal framework that is in place in order to provide individuals with a legal basis, which will protect their personal data privacy in the Big Data era.

5.1. CONCLUSION

Chapter 1 of the thesis consists of the introduction, Chapter 2 examines the technologies that are used for data analytics, such as HDFS and Map-Reduce. Subsequently, there is a discussion about whether the technologies to protect privacy are strong enough to fulfil their purpose as the Big Data analytics processes continue to grow exponentially. Part 2 of Chapter 2 analyses the legal principles and framework that exists, in order to protect individuals' personal data. There is a particular emphasis on how the General Data Protection Regulation is going to protect individuals' personal data, and the importance of the legitimate interest condition, the concept of consent and the data minimisation principle.

Chapter 3 identified the appropriate quantitative and qualitative research methods that were used during the course of the research process.

Chapter 4 provided a breakdown of the results and analysis of the surveys and the interviews.

There are doubts about whether the current privacy technologies are good enough to protect data privacy, for instance the Hadoop system is not very efficient, because when it is analysing and processing data it uses a parallel processing system to process data, which results in duplication of data. Also, it is an open source software system, which can result in a reduction in quality of the system functionality.

But recently, the security functionality of Hadoop has been given a lifeline in the form of project Rhino. This is an open source software which aims to provide greater support for encryption and key management. Furthermore, it includes a token based authentication framework and further strengthening of security auditing. (Smith, 2013)

The privacy preserving data mining process has proven to be very effective at protecting data privacy. Because it protects the data by altering it, which results in concealing or deleting the important and sensitive data to be masked. This method provides the means to resolve the data subject's original data from the altered data.

However, the PPDM is not without its faults also, as the data mining searches can result in scalability and overhead performance issues. Furthermore, the data is not anonymised as well as it should be in order to properly protect data privacy.

The main talking point from a legal perspective was the legal framework that is required in order to protect individuals' personal data privacy. This in particular refers to the new GDPR, which has ensured that the purpose limitation principle and the data minimisation principle are centre stage. While these principles will help in the regulation's determined efforts to protect data privacy, there are still some doubts in relation to secondary data use and consent.

As under the GDPR, consent means a "freely given, specific, informed and unambiguous indication of the data subject's wishes" signifying agreement to the processing of personal data relating to him or her. These requirements are unlikely to be satisfied if there is ambiguity and uncertainty around secondary processing purposes, as is often the case with data analytics where pre-obtained data sets are repurposed to unlock yet unrevealed value.

Consent or the other legal principles that may validate secondary processing should not be viewed in isolation. But, they should be considered by data controllers in the round with other principles by which accountability, the GDPR's focal principle, can be demonstrated.

The survey respondents' views and opinions revealed how they have a strong awareness of their privacy online, as they were aware of the importance of managing their privacy settings properly. They were also cautious with respect to their privacy online and do not approve of their data being used for purposes other than the original purpose. They want to have more control of their data, but yet use their real name online. They do so in order to take full advantage of the benefits of social networks.

It is possible to ensure that individuals can fully benefit from social media, while at the same time their data privacy can be protected. This involves the social media companies providing users with a simplified version of the rights and terms of use. Governmental departments should also ensure that its citizens are aware of their rights and ensure that social media companies comply with their obligations to these users, while working together to make the process easier.

Under the new GDPR, businesses must take other appropriate safeguards to ensure that the fundamental rights of data subject are upheld in those situations where the law regards that they take primacy over the data controller's interests. A novel approach to ensure that valid consent be obtained from data subjects online, could involve the incorporation of technological tools like personal data stores, blockchain, and other transparency enhancers, which give individuals the option of visualising their data use, and being able to follow their data after they change hands.

5.2. FUTURE RESEARCH

Looking at future research, there could be research and investigation into data management from a governmental point of view and the development of a government security framework. There should be a development of different security control mechanisms to match the respect security weak points.

If individuals' personal data was categorised, this would enable individuals to have an understanding of what is happening to their data. Furthermore, the GDPR obliges data controllers and organisations to properly inform data subjects of their rights.

Detailed data analysis could be carried out in the future, which could be achieved using prediction analysis and relationship modelling. Furthermore, graph databases could be used to analyse the metadata relationships between data and discovering the patterns in those relationships. Document databases are also very useful for storing unstructured data in a flexible way. (Burbank, 2016)

Cluster analysis can be used to identify subgroups of data with different patterns of scale scores. (Richardson, 2012)

Predictive analytics could also be used to extract data from existing data sets with the aim of identifying trends and patterns. Additionally, larger data sets can be used by the predictive analytics method, which can produce very reliable predictions based on the volume of data analyzed. (Galetto, 2018)

APPENDIX

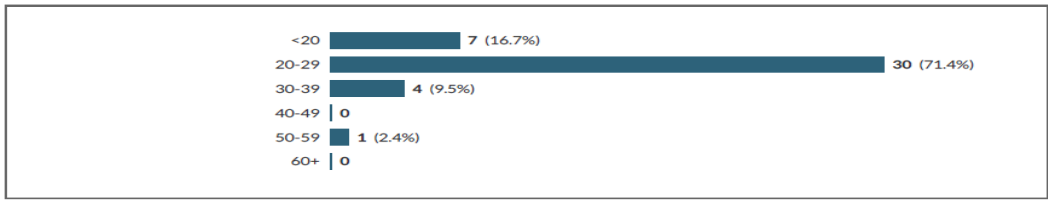
The following graphics are the images from the BOS survey.

bos

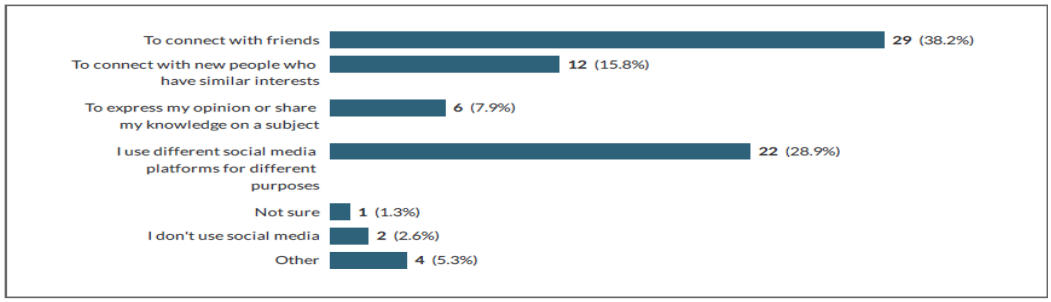
Social Media Use and Privacy Survey

Showing 42 of 42 responses
Showing **all** responses
Showing **all** questions
Response rate: 28%

1 Which of the following age groups do you belong to?



2 Why do you use social media?



2.a If you selected Other, please specify:

1 / 10

Showing all 4 responses	
none	265019-265011-22020947
Entertainment/news feed	265019-265011-22058106
To drive employee engagement through dialogue.	265019-265011-22077693
I use Google+ purely for Youtube	265019-265011-22166426

3 How do you manage the privacy settings of your social media accounts?

3.a If you selected Other, please specify:

No responses

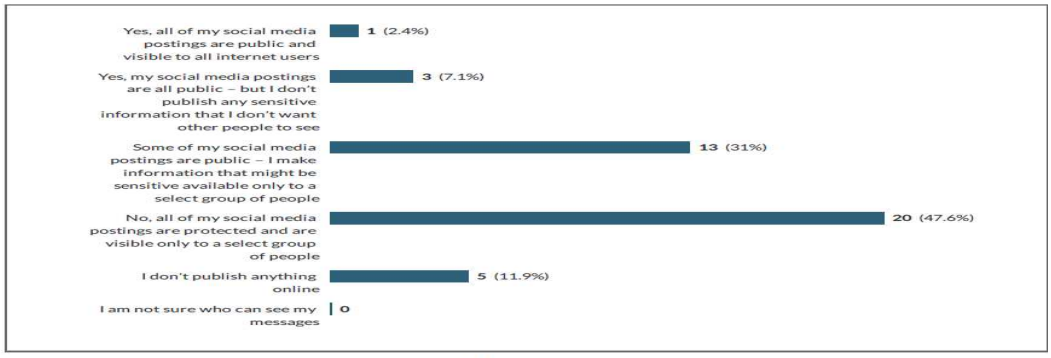
4 Do you make your personal contact information accessible to other internet users through your social media accounts?

2 / 10

4.a If you selected Other, please specify:

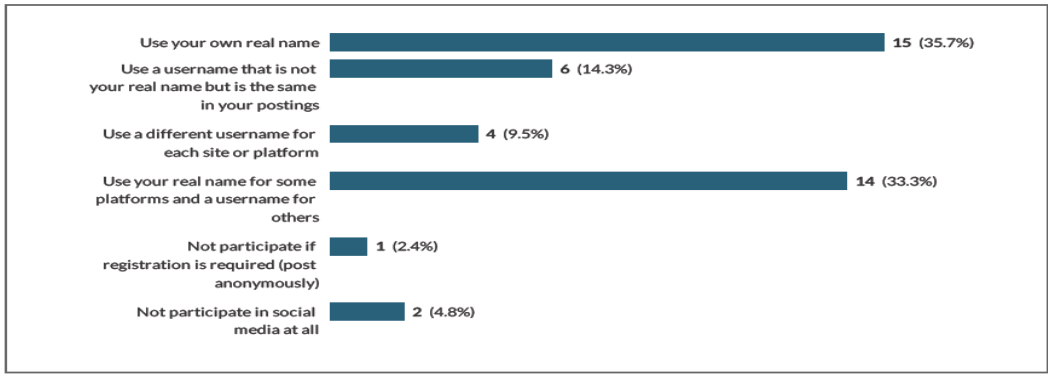
No responses

5 Are your Facebook and Twitter status update messages or blog posts you post in social media, visible to all internet users?



3 / 10

6 When you post things online in public venues (for e.g., Facebook messages or Tweets) do you prefer to:



7 Why do you use or not use your real name online?

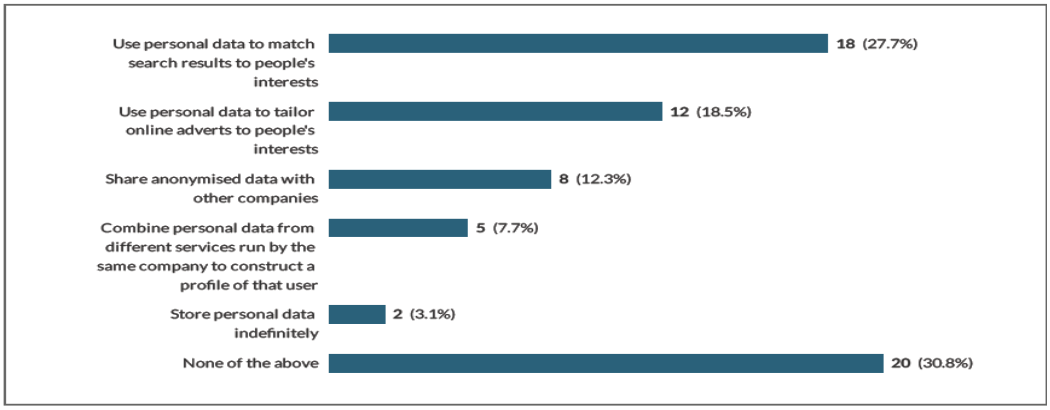
Showing all 42 responses	
Don't really want certain people to have that information	265019-265011-22020624
I like to protect my name I do not want anyone to be able to find me	265019-265011-22020835
Want to remain private	265019-265011-22020947
I feel no need to conceal my real name	265019-265011-22021707
N/A	265019-265011-22021955
So people can find me	265019-265011-22022970
So I look professional	265019-265011-22020497
Only use a personal Facebook profile - no need for other name.	265019-265011-22027992
Future employers	265019-265011-22031830
for privacy reasons	265019-265011-22047326
Use real name from habit when you had to - considering of changing now but haven't yet	265019-265011-22058106
I do not have the option to use a different name on social media	265019-265011-22064268
It's a variation of my real name so people who know me would know it's me	265019-265011-22075433
For some platforms, I prefer to have an identity people can recognise while still remaining anonymous or separated from that identity to anyone but the people closest to me.	265019-265011-22075372
I'd prefer to be given the choice, because sometimes my identity is not relevant to what is	265019-265011-22077693

4 / 10

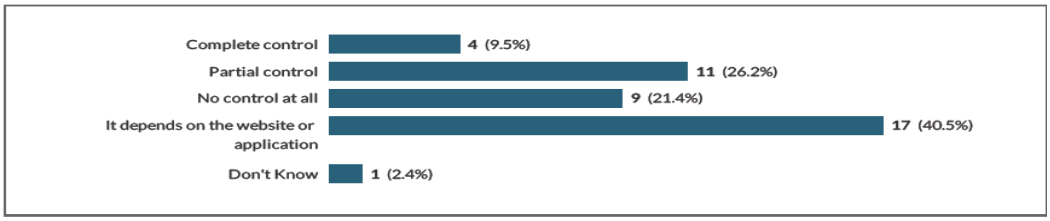
being posted and people don't need to know who I am in order to enter a dialogue with me. Increasingly though, more websites and removing the choice and forcing to use your real name.	
So that people know who it is posting	265019-265011-22077950
Oliver Bryant	265019-265011-22078481
I only use my real name for Facebook, and usernames for everything else	265019-265011-22079176
For some things I think it is more important to use my first name only	265019-265011-22091258
If i want to conceal my identity.	265019-265011-22094409
I like my privacy, theres the facebook version of me, then the reddit version of me which not everyone would understand	265019-265011-22094879
Bad habbit	265019-265011-22107056
When interacting with strangers on some platforms; using my real identity has no benefit but may open myself for identification	265019-265011-22113561
So it is harder to find me/know who I am	265019-265011-22113661
If my post is going to a wider audience I do not like any random person knowing my full name and using it to track me	265019-265011-22122548
I like being able to have a unique username, and some sites require it	265019-265011-22127717
Never thought to change it.	265019-265011-22139076
Generally don't post much on public venues, use my name on my private accounts.	265019-265011-22145453
because it is not safe on social media	265019-265011-22150745
Not afraid	265019-265011-22158144
I want people that I meet and know to be able to find me easily.	265019-265011-22162095
I use my IRL name on face book to make it easier for people to find me and for my friends to now it's me. On other platforms where I mostly talk to strangers I use a screen name for security reasons.	265019-265011-22163852
There's never been a need for it	265019-265011-22166426
Privacy concerns	265019-265011-22167784
So my friends know it is me when I try and connect with them, and when I don't use my real name it is protect myself so people don't find out personal information	265019-265011-22174507
Why is this a forced choice question? I've already said above I no longer have a social media account so this question does not really make sense for me to answer. If you are talking about emails etc. to it should be made clearer in your question.	265019-265011-22186513
Privacy reasons	265019-265011-22188812
so friends can find me	265019-265011-22191872
I don't have anything to hide, and I am transparent about my views and beliefs. I generally only use social media to post photos and connect with friends, rather than voice my opinion.	265019-265011-22239668
When I post, I post with the up,ost respect to society - however, for LinkedIn etc. I use my own name as to an extent you yourself are a brand, that at some point in the near future I (for	265019-265011-22241177

positive implications) I hope to be know within the field or subject area I wish to pursue	
So people I don't want to find me can't. Reduce chances of being stalked by people.	265019-265011-22288361
I use a username online to protect my online activity from potential employers.	265019-265011-22290799

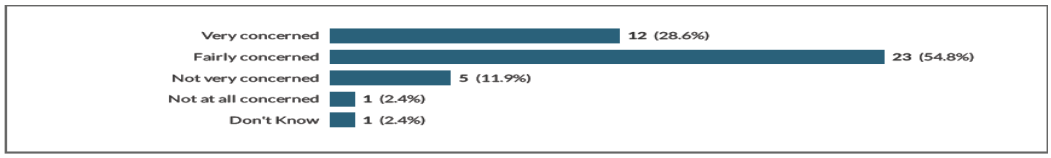
8 Which of the following, if any, do you think is acceptable for a company to do with online personal data (e.g. websites visited, products looked at, or emails sent)?



9 How much control do you believe you have over the information you submit online, e.g. the ability to correct, change or delete this information?

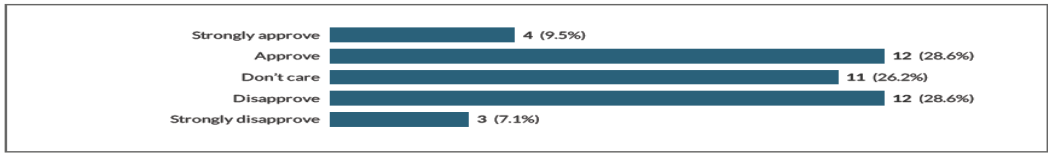


10 How concerned are you about not having complete control over the information you provide online? Would you say you are...?



Publicly Available Content and Privacy

11 Some companies collect, or "scrape," publicly accessible content that is posted on social media sites -- such as Twitter, blog and forum posts -- to discover what people say about different companies, brands and products. What do you think about the use of social media postings in this way?

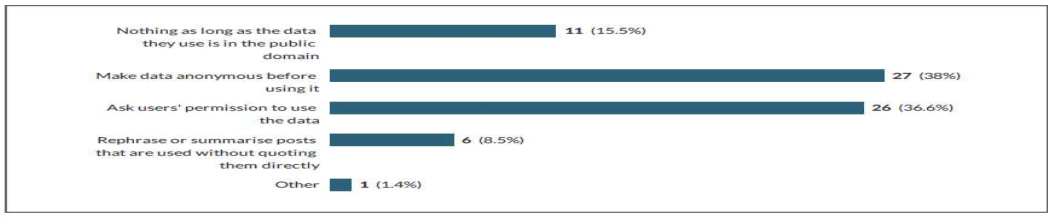


12 What worries do you have about companies scraping and analysing publicly accessible social media posts?

Showing 5 of 42 responses	
Don't really care	265019-265011-22020624
It can be beneficial but only for the actual company	265019-265011-22020835
Don't like it	265019-265011-22020947
None. If it's made public, what is he concern? It's public information	265019-265011-22021707
N/A	265019-265011-22021955

13 What must companies scraping for online data do to ensure social media privacy?

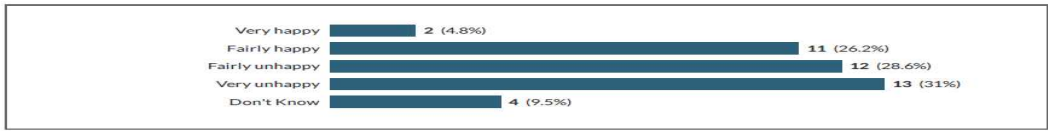
7 / 10



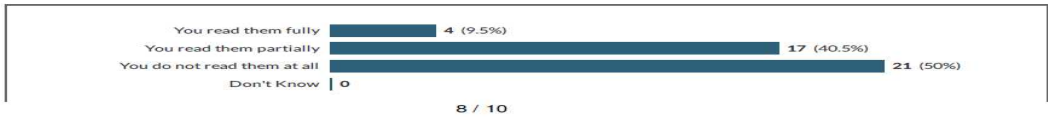
13.a If you selected Other, please specify:

Showing 1 response	
Stop scraping for online data	265019-265011-22058106

14 Some online companies are able to provide free services, such as search engines, e-mail accounts, etc., as a result of the income they receive from advertisers trying to attract users on their websites. How happy are you with the fact that those websites are using information about your online activity to mould or match, advertisements or content to your hobbies and interests?



15 Internet privacy statements describe how the personal information you submit will be used and who will have access to it. In relation to Internet privacy statements, which of the following best describes what you usually do?



8 / 10

Showing 5 of 42 responses	
Wasn't as aware of the concerns in this area until recently	265019-265011-22020624
I do not agree with the commercial use of social media data. I have been in situations before where certain websites have passed on my personal data to third party marketing companies. This is something I strongly disagree with as I do not wish to have marketing companies calling me with my name and address asking me to participate in something.	265019-265011-22020835
None	265019-265011-22020947
Everyone should be cautious about what they post on social media and ensure that privacy settings are in place when posting sensitive information.	265019-265011-22021707
I think the use of personal details to personalise advertising campaigns is what is drawing people to anonymity. If there was more control over how our information was used there wouldn't be so much need for secrecy measures online thus making criminal justice procedures easier	265019-265011-22021955

THE SURVEY MONKEY SURVEY QUESTIONS

The following 10 questions are those that were used in the two survey monkey surveys.

1. Why do you use social media?

- ☐ To connect with friends
- ☐ To connect with new people who have similar interests
- ☐ To express my opinion or share my knowledge on a subject
- ☐ I use different social media platforms for different purposes
- ☐ Not sure
- ☐ I don't use social media

Other views or opinions

2. How do you manage the privacy settings of your social media accounts?

- ☐ I change my privacy settings to control who has access to what I post on my account
- ☐ I am aware of different levels of privacy but don't worry about controlling them
- ☐ I am not sure how to manage the privacy settings on my social media accounts
- ☐ I don't use social media

Other views or opinions

3. Are your Facebook and Twitter status update messages or blog posts you post in social media, visible to all internet users?

- ☐ Yes, all of my social media postings are public and visible to all internet users
- ☐ Yes, my social media postings are all public – but I don't publish any sensitive information that I don't want other people to see
- ☐ Some of my social media postings are public – I make information that might be sensitive available only to a select group of people
- ☐ No, all of my social media postings are protected and are visible only to a select group of people
- ☐ I don't publish anything online
- ☐ I am not sure who can see my messages

4. Do you express your views on companies, products, services or brands through social media?

- ☐ Always when I have something to say
- ☐ Sometimes
- ☐ I have but only on a few occasions
- ☐ Have never done it but would consider it
- ☐ Have never done it and would not consider it

5. Some companies collect, or "scrape," publicly accessible content that is posted on social media sites -- such as Twitter, blog and forum posts -- to discover what people say about different companies, brands and products. What do you think about the use of social media postings in this way?

- ☐ Strongly approve
- ☐ Approve
- ☐ Don't care
- ☐ Disapprove
- ☐ Strongly disapprove

6. What worries do you have about companies scraping and analysing publicly accessible social media posts?

7. What must companies scraping for online data do to ensure social media privacy?

- ☐ Nothing as long as the data they use is in the public domain
- ☐ Make data anonymous before using it
- ☐ Ask users' permission to use the data
- ☐ Rephrase or summarise posts that are used without quoting them directly

Other views or opinions

8. Public bodies and private companies retaining information about you can sometimes use it for a different purpose than the one it was collected for, without informing you (e.g. for direct marketing, targeted online advertising, profiling). How concerned are you about this use of your information?

- ☐ Very concerned
- ☐ Fairly concerned
- ☐ Not very concerned
- ☐ Not at all concerned
- ☐ Don't Know

9. Do you believe that your explicit approval should be required before any sort of personal information is collected and processed?

- ☐ Yes, in all situations
- ☐ Yes, when personal information is required online
- ☐ Yes, when sensitive information whether online or offline is required (e.g. health, religion, political beliefs, sexual preferences, etc.)
- ☐ No
- ☐ Don't Know

10. Please express your thoughts and views, which you might have on the topic of online privacy in social media and the commercial use of public social media data.

BIBLIOGRAPHY

PRIMARY MATERIALS

Case Law

Griffin v. Board of Regents of Regency University, 795 F.2d 1281 (1986)

LEGISLATION

Section 48(1) and (2) of the Criminal Justice and Public Order Act 1994

UK LEGISLATION

Data Protection Act 1998

EU

DIRECTIVES

Council Directive No. 95/46 (Data Protection Regulation)

REGULATIONS

Commission Regulation 697/2016 (General Data Protection Regulation)

Commission Regulation 697/2016 (GDPR) Article 6(4) (c) (Lawfulness of data processing)

Commission Regulation 697/2016 (GDPR) Article 5(1) (c) (Data processing limited to what is necessary).

Commission Regulation 697/2016 (GDPR) Article 4 (Definitions).

Commission Regulation 697/2016 (GDPR) Article 9 (Processing of special categories of personal data).

SECONDARY MATERIALS

TEXTBOOKS

Baldwin, J. (1998) *Small Claims in the County Courts in England and Wales: The Bargain Basement of Civil Justice*. Oxford: Oxford University Press.

Beiske, B. (2007) *Research Methods. Uses and Limitations of Questionnaires, Interviews, and Case Studies.* Germany: GRIN Verlag.

Epstein, L and Martin, A. (2014) *An Introduction to empirical legal research*. Oxford, UK: Oxford University Press.

Hutchinson, T. (2006) *Researching and Writing in Law*. 2nd edn. Sydney, Australia: Thomson Publications.

Lankshear, C et al. (2004) *A handbook for teacher research.* Berkshire, UK: Open University Press.

Parise, E. (2011) *The Filter Bubble: What the Internet is Hiding from You*. New York: The Penguin Press.

Sachan, A et al. (2013) *An Analysis of Privacy Preservation Techniques in Data Mining.* In: Meghanathan, N et al. (eds) *Advances in Computing and Information Technology. Advances in Intelligent Systems and Computing.* Berlin, Heidelberg: Springer.

Schiffrin, D. (1994). *Approaches to Discourse*. Oxford and Massachusettes: Blackwell Publishers.

Silverman, D. (2011) *Qualitative Research, Issues of Theory, Method and Practice*. 3rd edn. London: Sage Publications.

Sokolova, M et al. (2015) *Personal privacy protection in time of big data. Challenges in Computational Statistics and Data Mining, Studies in Computational Intelligence.* Cham Berlin: Springer.

PERIODIC LITERATURE

Aberdour, M. (2007) Achieving Quality in Open Source Software. *IEEE Software,* 24(1), pp. 58-64.

Bethencourt, J et al. (2007) 'Ciphertext-policy attribute based encryption', *IEEE Computer Society Washington, DC,* pp. 321-334.

Bhala et al. (2016) 'Big Data Analytics for Social Network – The Base Study', *International Journal of Engineering Trends and Technology,* 36(9) pp. 467-470.

Boyd, W. (1972) 'Law in Computers and Computers in Law: A Lawyer's View of the State of the Art', *Arizona Law Review*, 14, pp.267-312.

Broeders, D et al. (2017) 'Big Data and security policies: Towards a framework for regulating the phases of analytics and use of Big Data', *Computer Law and Security Review,* 33(3) pp. 309-323.

Chen, J. (2016) 'How the best-laid plans go awry: the (unsolved) issues of applicable law in the General Data Protection Regulation', *International Data Privacy Law*, 6(4) pp. 310-323.

Chen, J et al. (2015) 'Secure transmission for big data based on nested sampling and coprime sampling with spectrum efficiency', *Security and Communications Networks,* 8(14) pp. 2447-2456.

Cheng, H et al. (2015) 'Secure big data storage and sharing scheme for cloud tenants', *China Communications,* 12(6) pp. 106-115.

Colombo, P et al. (2015) 'Privacy aware access control for big data: a research roadmap'. *Big Data Res*. 2(4) pp. 145–154.

Cradock, E. (2017) 'Nobody puts data in a corner? Why a new approach to categorising personal data is required for the obligation to inform' *Computer Law and Security Review,* 33(2) pp. 142-158.

De Hert et al. (2016) 'The new General Data Protection Regulation: Still a sound system for the protection of individuals?', *Computer law and Security Review,* 32(2) pp. 179-194.

Downs, R et al. (1997) 'If It Can't Be Lake Woebegone... A Nationwide Survey of Law School Grading and Grade Normalization Practices', *University of Missouri-Kansas City Law Review*, 65(4), pp.819-878.

Dubey, S. (2016) 'Implementation of Privacy Preserving Methods Using Hadoop Framework', *International Research Journal of Engineering and Technology,* 3(5) pp. 1268-1272.

Finch, E and Munro, V. (2008) 'Lifting the Veil: The Use of Focus Groups and Trial Simulations in Legal Research', *Journal of Law and Society*, 35, pp.30-51.

Gee, P. (1991) 'A Linguistic approach to narrative', *Journal of Narrative and Life History*, 1(1), pp.15-39.

Ghani, N et al. (2016) 'Big Data and Data Protection: Issues with Purpose Limitation Principle', *International Journal of Advances in Soft Computing and its Application,* 8(3) pp. 116-121.

Harris, L et al. (2010) 'Mixing interview and questionnaire methods:
Practical problems in aligning data', *Practical Assessment, Research & Evaluation,* 15(1) pp. 1-19.

He, S et al. (2016) 'Efficient group key management for secure big data in predictable large-scale networks', *Concurrency and Computation: Practice and Experience,* 28 pp. 1174-1192.

Henham, R. (2000) 'Reconciling Process and Policy: Sentence Discounts in the Magistrates' Courts', *Criminal Law Review,* 15, pp.436- 448.

Hurwitz, J. (1965) 'Three Delinquent Types: A Multivariate Analysis', *The Journal of Criminal Law, Criminology and Police Science,* 56(3), pp.328-334.

Irudayasamy, A et al. (2015) 'Scalable multidimensional anonymization algorithm over big data using map reduce on public cloud', *Journal of Theoretical and Applied Information Technology,* 74(2) pp. 221-231.

Jain, P et al. (2016) 'Big data privacy: a technological perspective and review', *Journal of Big Data,* 3(25) pp. 1-25.

Javed, Y et al. (2016) 'Defectiveness Evolution in Open Source Software Systems', *Procedia Computer Science,* 82 pp. 107 – 114.

Kitchin, R et al. (2016) 'What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets', *Big Data & Society,* pp. 1-10.

Law, S and Nolan J. (2003) 'Unintended and Persistent Consequences of Regulation: The Case of Cable Television Provision in Canada', *Journal of Network Industries*, 4(4), pp.391-410.

Lozowick, A et al. (1968) 'Law and Quantitative Multivariate Analysis: An Encounter', *Michigan Law Review*, 66, pp.1641-1678.

Luna, A. (2006) 'Faculty Salary Equity Cases: Combining Statistics with the Law', *The Journal of Higher Education,* 77(2), pp.193-224.

Lydia et al. (2016) 'Big Data Analytics: A Survey', *International Journal of Application or Innovation in Engineering & Management,* 5(10) pp. 90-106.

Maguire, E and Shin, Y. (2003) 'Structural change in large police agencies during the 1990s', *An International Journal of Police Strategies & Management,* 26(2), pp.251-275.

Maness, R and Valeriano, B. (2016) 'The Impact of Cyber Conflict on International Interactions', *Armed Forces & Society*, 42(2), pp.301-323.

Mehmood, A et al. (2016) 'Protection of Big Data Privacy', *IEEE Access,* 4 pp. 1821-1834.

Pence, H. (2015) 'WHAT IS BIG DATA AND WHY IS IT IMPORTANT?', *J. EDUCATIONAL TECHNOLOGY SYSTEMS,*43(2), pp. 159-171.

Qi, X et al. (2012) 'An Overview of Privacy Preserving Data Mining', *Procedia Environmental Sciences,* 12 pp. 1341-1347.

Rauhofer, J. (2013) 'One Step Forward, Two Steps Back? Critical observations on the proposed reform of the EU data protection framework', *Journal of Law and Economic Regulation*, 6(1) pp. 57-84.

Scaife, L et al. (2016) 'The GDPR and consent - A matter of child's play?', *Compliance and Risk,* 5(5) pp. 6-9.

Smithson J. (2000) 'Using and analysing focus groups: limitations and possibilities', *International Journal of Social Research Methodology*, 3(2) pp. 103-119.

Tabarrok, A and Helland, E. (1999) 'Court politics: The political economy of tort awards', *Journal of Law and Economics*, 42, pp.157-188.

Trinder, L and Stone, N. (1998) 'Family Assistance Orders – Professional Aspiration and Party Frustration', *Child and Family Law Quarterly*, 10(3), pp.291-302.

Ulbricht, L. (2016) 'Big data: big power shifts?' *Journal on Internet Regulation*, 5(1).

Wei, G et al. (2015) 'Obtain confidentiality or/and authenticity in Big Data by ID-based generalized signcryption', *Information Sciences,* 318 pp. 111-122.

Xiao, Z et al. (2013) 'Security and privacy in cloud computing', *IEEE Cloud Computing,* 15(2) pp. 845-859.

Xu, L et al. (2014) 'Information security in big data: Privacy and data mining', *IEEE Access,* 2 pp. 1149-1176.

Yoon, M et al. (2015) 'A data encryption scheme and GPU-based query processing algorithm for spatial data outsourcing', *BigComp,* 0 pp. 202-209.

Zheng, Y. (1998) 'Signcryption and its applications in efficient public key solutions', *Information Security*, 1396 pp. 291-312.

INTERNET RESOURCES AND REPORTS.

Abeyasekera, S. (2005) Statistical Services Centre, The University of Reading, *Multivariate methods for index construction.* Available at https://unstats.un.org/unsd/hhsurveys/finalpublication/ch18fin3.pdf. (Accessed July 2017)

Aldeen, S et al. (2015) SpringerPlus, *A comprehensive review on privacy preserving data mining.* Available at *https://springerplus.springeropen.com/track/pdf/10.1186/s40064-015-1481-x?site=springerplus.springeropen.com.* (Accessed June 2017)

Ballinger, C. (2011) *Why Inferential Statistics Are Inappropriate For Development Studies and How The Same Data Can Be Better Used.* Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1775002. (Accessed May 2017)

Bourka, A et al. (2015) European Union Agency for Network and Information Security, *Privacy by design in big data.* Available at https://www.enisa.europa.eu/publications/big-data-protection. (Accessed June 2017)

Boyen, X. (2006) IACR Cryptology, *Anonymous Hierarchical Identity-Based Encryption (Without Random Oracles).* Available at https://eprint.iacr.org/2006/085.pdf. (Accessed June 2017)

Burbank, D. (2016) Dataversity, *Data modelling for Big Data.* Available at: https://www.slideshare.net/Dataversity/data-modeling-for-big-data. (Accessed February 2018)

Buttarelli, G. (2016) European Data Protection Supervisor, *A smart approach: counteract the bias in artificial intelligence. European Data Protection Supervisor*. Available at https://edps.europa.eu/press-publications/press-news/blog/smart-approach-counteract-bias-artificial-intelligence_en. (Accessed April 2017)

Cavoukian, A. (2011) International Conference of Data Protection and Privacy Commissioners, *Privacy by Design.* Available at https://www.ipc.on.ca/wp-content/uploads/Resources/PbDReport.pdf. (Accessed July 2017)

Conboy, K. (2014) Irish Software Engineering Research Centre, *Embracing open innovation in agile software development.* Available at http://www.engineersjournal.ie/2014/02/13/embracing-open-innovation-in-agile-software-development/. (Accessed February 2018)

Custers, B et al. (2016) International Data Privacy Law, *Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection.* Available at http://data-reuse.eu/wp-content/uploads/2016/01/International-Data-Privacy-Law-2016-Custers.pdf. (Accessed June 2017)

Dallal, G. (2014) Jerrydallal.com, *Non Parametric Statistics.* Available at http://www.jerrydallal.com/lhsp/npar.htm. (Accessed December 2016)

Davies, R. (2016) European Parliamentary Research Service, *Big data and data analytics the potential for innovation and growth.* Available at http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/589801/EPRS_BRI(2016)589801_EN.pdf. (Accessed June 2017)

European Report Working Party (2014) Article 29 Data Protection Working Party, *Impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU.* Available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp221_en.pdf. (Accessed May 2017)

Figure 1: Table summarising the different methods used by researchers, with respect to qualitative and quantitative research method types., 2016. Available at: learn.maricopa.edu, 2016. (Accessed: 10 November 2016)

Galetto, M. (2018) *What is Predictive Analytics? Definition and Models.* Available at: https://www.ngdata.com/what-is-predictive-analytics/. (Accessed February 2018)

Gentry, C. (2009) Computing Dept, Stanford University, *A Fully Homomorphic Encryption Scheme.* Available at https://crypto.stanford.edu/craig/craig-thesis.pdf. (Accessed July 2017)

GU, C. (2014) Microsoft Blogs Tech Notes, *An overview of Hadoop Distributed File System, HCatalog, Hive and map-reduce.* Available at https://blogs.msdn.microsoft.com/technotes/2014/12/21/an-overview-of-hadoop-distributed-file-system-hcatalog-hive-and-map-reduce/. (Accessed June 2017)

Holman, M. (2016) Duke University Law Library, *A Brief Introduction to Empirical Legal Scholarship.* Available at: https://**law**.**duke**.edu/lib/downloads/ELS**Intro**.ppt (Accessed: 1 November 2016)

Information Commissioner's Office. (2016) Information Commissioner's Office, *Big data, artificial intelligence, machine learning and data protection.* Available at https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf. (Accessed March 2017)

Irish Data Protection Commissioner. (2017) Irish Data Protection Commissioner Documents, *Article 29 Working Party.* Available at https://www.dataprotection.ie/docs/Article-29-Working-Party/181.htm. (Accessed May 2017)

Jarvis, L and Macdonald, S. (2015) 'What Is Cyberterrorism? Findings From a Survey of Researchers, Terrorism and Political Violence', 27(4), pp.657-678. doi: 10.1080/09546553.2013.847827.

Kepner, J et al. (2014) High Performance Extreme Computing Conference, *Computing on masked data: A high performance method for improving big data veracity.* pp. 1-6. Available at https://arxiv.org/ftp/arxiv/papers/1406/1406.5751.pdf. (Accessed June 2017)

Kromhout, M. and van San, M. (2003) 'Shadowy worlds. New ethnic groups and juvenile, Cashier', The Hague: WODC. Available at: http://repository.tudelft.nl/view/wodc/uuid%3Af1e83efb-d3dd-4e86-862a-c1abc2fd009a (Accessed: 1 November 2016)

Mason Hayes and Curran. (2017) Mason Hayes and Curran Tech Law Blog, *Getting Ready for the General Data Protection Regulation, A Guide by Mason Hayes & Curran.* Available at https://iabireland.ie/getting-ready-for-the-gdpr-guide/. (Accessed May 2017)

Moreno, J. (2016) Future Internet, *Main Issues in Big Data Security,* 8(44) pp. 1-16. Available at http://www.mdpi.com/1999-5903/8/3/44. (Accessed June 2017)

Oostveen, M. (2016) International Data Privacy Law, *Identifiability and the Applicability of Data Protection to Big Data.* Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2877692. (Accessed April 2017)

Ostrovskii, D. (2013) Chalmers University of Technology, *Trend analysis in input data for PSA.* Available at http://studentarbeten.chalmers.se/publication/180158-trend-analysis-in-input-data-for-psa. Accessed (June 2017)

Paraccel. (2012) Paraccel, *Hadoop's Limitations for Big Data Analytics whitepaper.* Available at https://www.whitepapers.em360tech.com/wp-content/files_mf/1360922634PARACCEL1.pdf. (Accessed July 2017)

Quan, Z et al. (2013) International Conference on Emerging Intelligent Data andWeb Technologies, *Trusted Scheme for Hadoop Cluster.* pp. 344-349. Available at http://dl.acm.org/citation.cfm?id=2547574. (Accessed July 2017)

Richardson , J. (2012) The Open University, *Quantitative data analysis: Planning an analytic strategy for your research.* Available at https://www.slideshare.net/OUmethods/research-methods-john-richardson?qid=c1a763e1-9cb9-4c19-a89a-55b9e1785009&v=&b=&from_search=10. (Accessed February 2018)

Savvides, S et al. (2014) International conference on automated software engineering, *Program analysis for secure big data processing,* pp. 277-288. Available at https://www.cs.purdue.edu/homes/ssavvide/publications/spr.pdf. (Accessed June 2017)

Smith, K. (2013) Big Data Security: The Evolution of Hadoop's Security Model. Available at https://www.infoq.com/articles/HadoopSecurityModel. (Accessed September 2017)

Techopedia, (2017) Techopedia, *Definition - What does NoSQL mean?* Available at https://www.techopedia.com/definition/27689/nosql-database. (Accessed June 2017)

Techtarget, (2017) Techtarget, *Definition on Hadoop Distributed File System.* Available at http://searchbusinessanalytics.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS. (Accessed June 2017)

Techtarget, (2017) Techtarget, *Definition on Apache Hive.* Available at http://searchdatamanagement.techtarget.com/definition/Apache-Hive. (Accessed at June 2017)

Thuraisingham, B. (2015) Codaspy'15, *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy.* Available at http://dl.acm.org/citation.cfm?id=2699026. (Accessed June 2017)

Tutorialspoint. (2017) Tutorialspoint, *Tutorial on Hadoop.* Available at https://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm. (Accessed June 2017)

Tutorialspoint. (2017) Tutorialspoint, *Tutorial on Hadoop mapreduce.* Available at https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm. (Accessed June 2017)

Ulusoy, H et al. (2015) ACM Symposium on Information, Computer and Communications Security, *Fine-grained Security Policy Enforcement for MapReduce Systems,* pp. 285-296. Available at http://dl.acm.org/citation.cfm?id=2714624. (Accessed June 2017)

Willemsen, M. (2016) University of Twente, *Anonymizing Unstructured Data to Prevent Privacy Leaks during Data Mining.* Available at http://referaat.cs.utwente.nl/conference/25/paper/7547/anonymizing-unstructured-data-to-prevent-privacy-leaks-during-data-mining.pdf. (Accessed July 2017)

World Economic Forum. (2013) World Economic Forum, Industry Agenda, *Unlocking the Value of Personal Data: From Collection to Usage.* Available at http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf. (Accessed May 2017)

Xu, Z et al. (2011) International Conference on Digital Information Management, *Classification of privacy-preserving distributed data mining protocols*. Available at http://ieeexplore.ieee.org/document/6093356/. (Accessed June 2017)