Deniz Anttila

# BIG DATA AND PRIVACY IN THE TELECOM INDUSTRY

JYVÄSKYLÄN YLIOPISTO
TIETOJENKÄSITTELYTIETEIDEN LAITOS
2019

# TIIVISTELMÄ

Anttila, Deniz
Big data ja yksityisyys teleoperaattorialalla
Jyväskylä: Jyväskylän yliopisto, 2019, 57 s.
Tietojärjestelmätiede, pro gradu -tutkielma
Ohjaaja: Seppänen, Ville

Teleoperaattorien tuottamien verkkoyhteyksien kautta kulkevan tietoliikenteen määrä on kasvussa johtuen tietoliikenteen saatavuuteen nojaavista digitaalisista palveluista, kuten television ja musiikin suoratoistopalveluiden käytön kasvusta.

Teleoperaattorit voivat hyötyä liikenteen kasvun määrästä, sillä operaattorit ovat ainutlaatuisessa asemassa hyödyntämään asiakkaidensa tuottamaa dataa. Big data teknologian käyttö tarjoaa useaa kaupallista hyötyä teleoperaattoreille, kuten asiakasvaihtuvuuden minimointi, tietolähteitä parempien strategisten päätösten tekoon, sekä parempaa mahdollisuutta palveluidensa laadun tarkkailuun.

Asiakkaiden tuottaman datan hyödyntäminen voi olla teknisesti hankalaa, mutta voi myös tuottaa asiakkaiden keskuudessa mielipahaa. Asiakkaat saattavat kokea datan hyödyntämisen loukkaavan yksityisyyttänsä, esimerkiksi jaettaessa tietoja kolmannen osapuolen kanssa. Tämä saattaa johtaa asiakkaiden katoon tietoliikennedataa hyödyntävistä teleoperaattoreista. Tämän pro gradu –tutkielman tavoitteena on tutkia hyötyjä ja haittoja koskien teleoperaattoreiden trendiä käyttää tietoliikenteestä poimittua dataa kaupallisesti hyödyksi big data teknologian avulla.

Tutkielman teoreettinen osuus perustuu kirjallisuuskatsaukseen aiheen aikaisemmista tutkimuksista. Tutkielman tuloksena on selkeämpi näkemys siitä, mitä teleoperaattorilta edellytetään asiakkaan yksityisyyden suhteen hyödynnettäessä big data teknologiaa ja kuinka teleoperaattorien asiakaskunta kokee datan hyödyntämisen.

Tutkielman empiirisessä osuudessa tarkastellaan valittua teoriamallia kyselystä kertyneiden vastauksien avulla. Osuuden tuloksia verrataan muihin tutkimustuloksiin.

Asiasanat: teleoperaattoriala, big data, yksityisyys.

# ABSTRACT

Anttila, Deniz
Big data and privacy in the telecom industry
Jyväskylä: University of Jyväskylä, 2019, 57 p.
Information Systems, Master's Thesis
Supervisor: Seppänen, Ville

While the telecommunications industry is digitalizing, the data traffic of broadband and mobile subscriptions is growing due to new digital services like streaming media.

The telecom industry is finding new ways of keeping the business of upgrading the network appropriately attractive. The issue is not limited to the technical side of handling the growing traffic in the network, but the business side on how to transform that traffic profitable for the companies who roll out the network. Communication service providers make sure that the services they offer are generating revenue by implementing big data technology. Service providers are in a unique position when handling data generated by their customers and big data offers number of commercial benefits as well as benefits in service quality.

Using customer generated data for a profit can be an issue, since the customer might perceive that his or her privacy is being violated when their data is shared to a third party. This might lead to increase of customer churn for the service providers which use data gathered from their customers. This thesis paper aims to study the pros and cons of the trend of big data implementation in the telecom industry.

The theoretical part for this thesis will be performed by a literary review on former research in this field. In the end the results of this thesis drive to create a clearer view of what is to be expected from a service provider on customer privacy issues when implementing big data and how do the customers perceive it.

In the empirical part of this thesis a selected theoretical model will evaluated through gathered responses from an online survey. The results are then compared to other results in the field.

Keywords: telecom, big data, privacy.

## FIGURES

## TABLES

## ATTACHMENTS

# TABLE OF CONTENTS

# 1 INTRODUCTION

The ability to organize and analyze vast amounts of data from several data sources is becoming more important as the amount of data generated daily is increasing globally. Increase in the usage of smart devices as well as the popularity of social networking applications are generating a growing flow of data that can reveal customer demographics, spending behavior, lifestyle and social influences (Van Den Dam, 2013). It is estimated that the amount of data is predicted to increase 300 times from 2005 to 2020 (Matturdi, Xianwei, Shuai & Fuhong, 2014). To put the growth of the generated data to perspective, the years 2013 and 2014 generated roughly 90 percent of the world's data in the year 2015 (O'Leary, 2015).

Large amounts of data indicate that there is a large amount of information that can be gained through analyzing the data. This information can be used to make more educated decisions and more likely predictions in the private sector as well as in the public sector. Big data and its techniques are becoming more common to address the issue of utilizing and analyzing data. The term "big data" refers to a large collection of information, typically stemming from more than one data source, and eventually being processed by a data processor or data analyst (Jensen, 2013). The term "big data" will be explored more in the first section of this paper.

Van Den Dam (2013) concluded that especially the communication service providers (CSPs) have a unique position to access the wealth of data of their customers, which no other industry has, and utilizing the data might eventually become a core capability. The CSPs have large subscriber bases, and data is generated every time a customer makes a call, sends a text or uses the internet. In addition, smartphones are basically on all the time and provide the CSPs constantly with new data, which makes it possible for the CSPs to know their customers better than most other industries. Basically, this means that due to the emergence of smartphones, tablets and other devices that are application dependent, the volume of signaling data (i.e. non-message information about the device, its location and updates) has also increased. The CSPs are increasingly seeing the potential of big data as more than half (54 %) of the CSPs where in the process of developing a strategy roadmap on big data and how to apply it to business challenges in 2013 (Van Den Dam, 2013.)

Some of the data gathered can be considered as personal information, which gives rise to the issue of information privacy in the techniques in which the data is analyzed. Big data technologies may affect the meaning of such concepts as data ownership, privacy, and identity for both organizations and individuals as information is aggregated and correlated by not only the originating, but also by those who may seek to further use and change it. It is hardly controlled how this information is used once it is out of the CSPs hands (Voronova, 2015.)

According to O'Leary (2015) privacy refers to the ability of an individual to maintain or ensure information asymmetries over that individual's data and limit others from knowing or learning about them. Security and governance of personal information is becoming even more important and daunting as CSPs embrace new sources of information, especially social media data (Van Den Dam, 2013). Then again, in favor of the implementation of big data in a telecommunication company, Arora & Malik (2015) explained that customer privacy might be even enhanced, since it is expected that big data analytics will impact various aspects of information security such as network monitoring, user authentication control, authorization, identity management, fraud detection, data loss prevention and data control.

This thesis is motivated by the issue that the implementation and utilization of big data techniques is becoming the norm for CSPs in their effort to capitalize on customer generated data. Since the data gathered by the CSPs can have characteristics of being personal information about their own customers, it is necessary to research the impact that this trend might have on the privacy of customer data. Additionally, the customers' attitudes towards CSPs implementing big data techniques on data generated through the customer's online activity, or even activity in general, is a worthy subject to do research on. Furthermore, the attitudes of the customers towards this change is important, since negative attitude might lead to increase in customer churn and negative business impact for the CSPs. Customer churn is a term used in the telecommunication service industry to describe the customer movement from one service provider to another (Kamalraj & Malathi, 2013). In turn, positive attitudes by providing greater customer experience might increase customer loyalty and limit customer churn (Van Den Dam, 2013). This thesis drives to understand and outline the positive as well as the negative issues of which the implementation of big data techniques in CSPs might disclose.

## 1.1 Key terms

The key terms of this thesis are associated with technical terms used when referring to data management technologies, business terms of the communication service industry and terms concerning data privacy. The most significant and likely the most indecisive term is "big data". There has been much discussion on the meaning of big data, but the following definition was selected.

| **Term** | **Definition** |
|---|---|
| **Big Data** | A data environment in which scalable architectures support the requirements of analytical and other applications which process, with high velocity, high volume data which may have a variety of data formats and which may include high velocity data acquisition. |
| **CSP** | Communication Service Provider.<br>Company that provides telephone subscription services for private consumers. |
| **Telecom** | Telecommunication business |
| **Churn** | Term used in the telecommunication service industry to describe the customer movement from one service provider to another |

**Table 1** Terms & Definitions

## 1.2 Research objective, question and limitations

The utilization of big data by telecommunication service providers has not been studied extensively. However, Libaque-Sáenz, Wong, Chang, Ha & Park (2016) sought out to create a model of perceived information risks that communication service subscribers have towards their CSPs when using customer data for additional benefits. They observed that information practices play an important role in shaping information privacy concerns, trust, and perceived information risks by surveying panel members on the internet in Korea. Libaque-Sáenz et al. (2016) concluded that the CSPs will be able to benefit from the large amount of data processed by their servers only when their customers see lesser risks in allowing CSPs to use their personal information.

Libaque-Sáenz et al. (2016) proposed to study further the differences in the sensitiveness towards personal information in different cultures and how it affects their willingness to share information with CSPs. Some cultures might perceive a piece of information as highly personal while in other cultures it might be seen as merely public information. In support of studying the subject in Finland, Statistics Finland (2016) identified that most of the big data technology development has been in the business sector of information and communications.

Thus, previous research shows that big data is becoming more utilized in the telecommunication sector in Finland and that data privacy plays an important role in the trend. The objective of this study is to elucidate the negative and positive privacy related issues that the implementation of big data techniques might have in the telecommunication sector. This study deals with factors which might be considered affecting the privacy of individuals who subscribe to mobile services. The following research question is set for the examination of the research problem:

- What are the privacy related issues of implementing big data techniques in telecommunication companies?
- How do communication service subscribers consent to their data being analyzed?

By providing a credible study relating to the research questions, it is the intention that this thesis can produce support for telecommunication companies in implementing big data techniques in a rational manner. The objective is to highlight aspects which should be taken into consideration by the CSPs for the positive results in doing so would be maximized, while in contrast the negative effects would be minimized.

This thesis will not drive to explain and analyze the ethical aspect of utilizing big data in telecommunication industry. Explaining ethics is more of a subjective matter which depends on individual perceptions on what is to be considered right or wrong. More specifically this thesis aims to observe the business impact a telecommunication service provider might experience due to privacy related issues of big data utilization. However, it cannot be fully excluded that the findings of this study might uncover some ethical perceptions of the subscribers of communication services.

This thesis will proceed in the second chapter by first reviewing the term of big data and how it is defined in the academic literature. Next, the possible positive and negative business-related outcomes of implementing big data in telecommunication companies to gain economical gains are explored. The third chapter focuses on what are regarded as privacy related matters. The privacy regulations set by the public authority and the consequences they produce are explored as well. Factors, which have been seen to be important in enhancing the privacy in the implementation of big data in a telecommunication company, are reviewed on the last subchapter. Lastly, a summary of this study is conducted at the last chapter of this thesis. There the meaning and the constraints of this paper and the emerged subjects for further research are reflected upon.

# 2 THEORETICAL FRAMEWORK

## 2.1 Previous research and literature

International scientific journals and conference publications are key publication channels for scientific information in the field of information technology. This study uses literature from foreign and some domestic studies, textbooks, articles, journals and conference publications. The theoretical part of this thesis is a literature review of academic publications gathered from source libraries which have been recognized as significant in information system research (i.e. European Journal of Information Systems, Information Systems Journal & Institute of Electrical and Electronics Engineers).

When searching for literature for this study, topics referring to big data, privacy and the combination of the former two relating to the telecommunication industry were the key search criteria (key search words: big data, privacy, telecom, & communication service provider). The literature was searched mainly through the websites of the different journal databases. Because most of the literature was searched through web portals, it is evident that physically printed literature (e.g. textbooks in libraries) were neglected when writing this thesis. Majority of the material reviewed in this study's literature review has been collected from the digital libraries of the Institute of Electrical and Electronics Engineers (i.e. IEEE Xplore Digital Library) and the Association for Computing Machinery (i.e. ACM DL).

The purpose of the literature review is to provide perspective into the current situation regarding the research of big data utilization in the telecommunication industry and what relevant privacy related issues have been noted. The term "big data" itself has discrepancies in the current literature and they will be outlined in the next chapter while defining the term.

In order to indicate the significance and impact of the studies used in this thesis, Google Scholar was used to review the amounts of citations done towards the studies. It is arguable that the amount of citations done on a study depicts the significance of it. Yet, newer studies have less citations than older ones. In general, the purpose of the theoretical research is to study academic publications in effort to find the most common issues in big data privacy from the perspective of telecommunication companies.

Big data has been studied in detail during the past decade as the realization of ever-growing data amounts generated daily has become more common. The utilization of big data in telecommunication companies and the privacy related issues which might emerge have been studied more lightly during the past decade. Nevertheless, general privacy issues of big data have been studied extensively and their results can often be applicable to the issues faced by the CSPs.

Libaque-Sáenz et al. (2016) studied factors which influence customers' risk perception on information risks regarding giving the telecommunication companies permissions for the usage of their personal information. The study was performed by a survey of 512 internet panel members in Korea. The goal of the study was to develop a model of perceived information risks and validate it with data from the telecommunications sector. The model was to serve as a framework for CSPs in formulating effective strategies for managing customer risk perception. The theoretical model of Libaque-Sáenz et al. (2016) is also explored in the empirical research of this thesis.

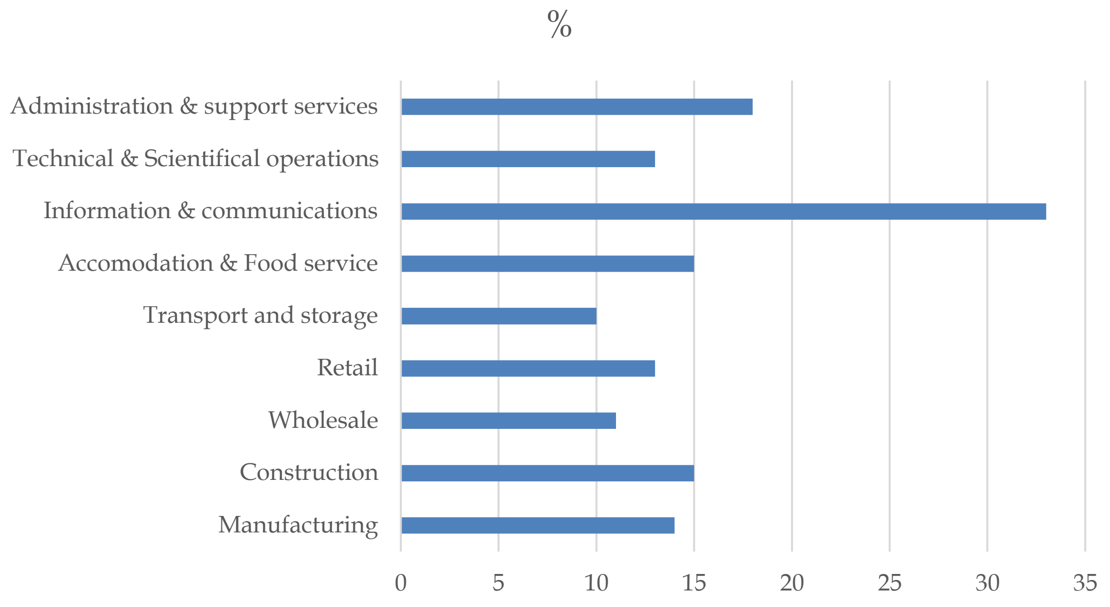

**Figure 1** Big data usage in Finland (Statistics Finland, 2016)

In relation of big data use in Finland, Statistics Finland conducted a survey in the spring of 2016 which was sent out to 4329 different companies to explore the use of information technology in those companies. 2981 (69%) of the companies replied to the survey in an acceptable manner. The companies represented comprehensively the whole spectrum of different industry sectors and company sizes in Finland. The survey included questions about the situation of the utilization of big data in the companies. The findings were that 15% of all the survey respondents had utilized big data and most of which (33%) were in the industry of information and communications (Figure 1). The finding highlights the relevance of big data utilization especially with telecommunication service providers. The term of "big data" was acknowledged not to be unambiguous and even with an added definition some uncertainty was expected in the statistics (Statistics Finland, 2016.). The following chapter will strive to explain and define big data in more detail.

## 2.2 Big Data

The world population is generating data in an ever-exponential rate by the help of technological advancements in data networks, data processing and storage capabilities in the recent decades. Companies in turn gather data from various sources i.e. from purchases done by credit cards, search terms used on Google, or even locational information gathered from mobile phones. Not only are the technological advancements the cause for the vast amounts of data generated, but the culture of western society has changed as technology has changed. Due to digital services like social media platforms (i.e. Facebook, Instagram, Snapchat etc.) and the cultural change that they have given rise to, people are willing to share pictures, videos, and statuses of themselves more often publicly through these services (Debatin et al, 2009). The daily generated data which an individual person generates is versatile in its nature and can provide substantial insight in demographical behavior. This requires tools which can efficiently analyze the data, also referred to as "big data".

### 2.2.1 Definition

There are as many definitions for big data in the information system literature as there are authors writing about it. In order to clarify the discussion about the definition of big data it is essential to categories different types of definitions. Hu, Wen, Chua & Li (2014) outlined three types of definitions which play a role in how big data is viewed. The types outlined where *attributive definition, comparative definition* and *architectural definition*.

An attributive definition depicts the features of big data, such as the volume, variety, velocity, veracity and value; also known as the "5Vs" definition. According to the definition volume refers to the large amount of data that is stored and analyzed. Velocity is referring to the high speed of data in which it is generated and should be processed for gathering insight. The more data is processed and analyzed in real time, the more value it will provide. Variety indicating heterogeneous data types and sources which make some of the gathered data unstructured and not applicable with typical data formats. Veracity of data then describes the consistency and reliability of it. When data is gathered from more sources and supported in wider forms the quality and accuracy of it is improved as well. Finally, value refers to the economic benefits gained by analyzing the data and utilizing the insights it holds (Terzi, Terzi & Sagiroglu, 2015; O'Leary, 2015; Hu et al., 2014.).

As presented above, there are several papers arguing to expand the original 3Vs definition (volume, velocity and variety) described on Doug Laney's 2001 research report about the data growth challenges and opportunities. The report was intended to highlight the three-dimensional challenges and opportunities of increasing volume, velocity and variety of data, but ended up being used as a definition for big data and later sought to be expanded further.

A comparative definition of big data is a subjective and non-metric definition. This type of definition does not bind itself to any specific metrics e.g. the amount of data in terabits and therefore can be considered a relevant definition even in the future keeping in mind the likely technical advances in the field. In other words, the definition type incorporates an evolutionary aspect of what is required of a dataset for it to be considered as big data (Hu et al., 2014).

It is almost impossible to try to define the amount of data that a dataset needs to have for it to be considered "big data". E.g., defining big data in terms of a dataset exceeding a specific number of terabytes is unnecessary, since it can be assumed that technology will have advances over time and the size of the datasets, which can be qualified as "big data", will most likely increase. The amount of data also varies between different industries as well as depending on the software and tools and what size of datasets are common in each particular industry (Manyika et al., 2011.).

Manyika et al. (2011) gave a comparative definition to big data by stating that it commonly refers to datasets which size is beyond the ability for typical database software tools to capture, store, manage, and analyze. The downside of defining big data this way is that it is generic and lacks the description of the usability and value of the data.

An architectural definition drives to further categorize big data into big data science and big data frameworks. Big data science studies the techniques in which big data is acquired, managed and evaluated. Big data frameworks refer to software libraries with algorithms which enable distributed processing and analysis of big data (Hu et al., 2014).

Some would consider it to be logical to embrace all alternative definitions of big data (Hu et al., 2014), but some argue that the variety of definitions disclose the need for a shared understanding of big data to avoid inconsistency and duplication (Emmanuel & Stanier, 2016). After investigating several definitions in the information system literature this paper will hold the following definition provided by Emmanuel & Stanier in their paper "Defining Big Data" (2016) as the basis depiction of big data:

> The term Big Data describes a data environment in which scalable architectures support the requirements of analytical and other applications which process, with high velocity, high volume data which may have a variety of data formats and which may include high velocity data acquisition. (Emmanuel & Stanier, 2016).

The definition above can be identified to include characteristics of all the definition types described earlier in this chapter and hence its relevance will not be expected to deteriorate due to technological advancements.

Additionally, to the definition given above, this thesis specifically refers to big data as the large datasets generated by communication service subscribers through their usage of different digital devices (e.g. mobile phones). Furthermore, this paper focuses on the analyzation of big data by referring to the extraction of hidden insight about consumer behavior and the exploitation of that insight through advantageous interpretation (Erevelles, Fukawa & Swayne, 2016). Advantageous interpretation includes the distribution and selling of consumer behavioral insight to third party members. The third-party members then are able to utilize the insight for better decision making when aiming to promote, market and sell their own products or services.

### 2.2.2 Data types

Big data consists of different data types due to the many sources from which the data is gathered from. As the data is gathered in variety of formats special techniques are required to utilize it. The variety of the data introduces new types of data, or non-traditional data, which have not been traditionally analyzed. Dhar & Mazumbar (2014) discussed about the key challenges in utilizing traditional and non-traditional data under the same data management strategy. Overall, they classify the generated data volume to three categories; transactional data, observational data and social data or interaction data.

Transactional data is usually generated by controlled interactions between enterprise and its customers and other internal/external stakeholders through a defined set of enterprise applications and interfaces. The structure of this data is designed and is relational in its nature. This data can be generated from e.g. ecommerce applications. A CSP might gather transactional data from its subscribers when they order their mobile subscription for the first time and continue to pay their subscription bills. Due to smartphones being used for purchases (e.g. mobile payment or recharging a prepaid account), the term "extended data record (XDR) has been created to describe purchases made in various ways (Van Den Dam, 2013).

Observational data is generated by machines or sensors as ancillary to the main application data while business processes get executed. Observational data is non-relational in nature though usually it has been decided in its design phase. This data might refer to data generated by sensors that monitor events or even customer call logs in call centers.

Social data or interaction data has typically open or free flowing structure and it has not been decided in the design time of other processes or applications.

Social data can be gathered through defined process or casual interactions like customer feedback or information gathered from social media (e.g. FaceBook, Twitter and LinkedIn).

Other data types can include signaling data. Signaling data is non-message information about the device, its location and updates. This data is transmitted from the subscribers' mobile device to the CSP even if the subscriber is not using it.

When the different data types are brought together for analyzing, it may result in contextual data about the individual. Contextual data is data which gives context to a person while taken from various sources e.g. smartphones, tablets, personal computers, networks, sensors, RFID tags and social media (Van den Dam, 2013). It may include information of the individual's family, socioeconomic background, general environment, educational history or even health background. Some of the contextual data may be regarded as highly private data which the individual does not wish to be known by any other person.

While Voronova (2015) was researching the ethical aspects of big data he used the formula of Russel Ackoff who in 1989 classified the content of the human mind into five categories: wisdom, and below understanding, knowledge, information and data (Figure 2). In the model data in itself does not have value until it is processed into a usable form in order to become information. As data is processed higher up in the hierarchy, it becomes more valuable and eventually it will be seen as wisdom. Ackoff indicated that the first four categories of the hierarchy relate to the past and they deal with what is already known and what has happened. Only the top of the hierarchy - wisdom is seen as dealing with the future rather than just grasping ideas about the present and past (Voronova, 2015). The objective of big data technologies and techniques is to process available data as efficiently as possible to produce wisdom in the form of predicting future trends. One of the main features of big data technologies and techniques is to identify patterns and perform predictions out of them.

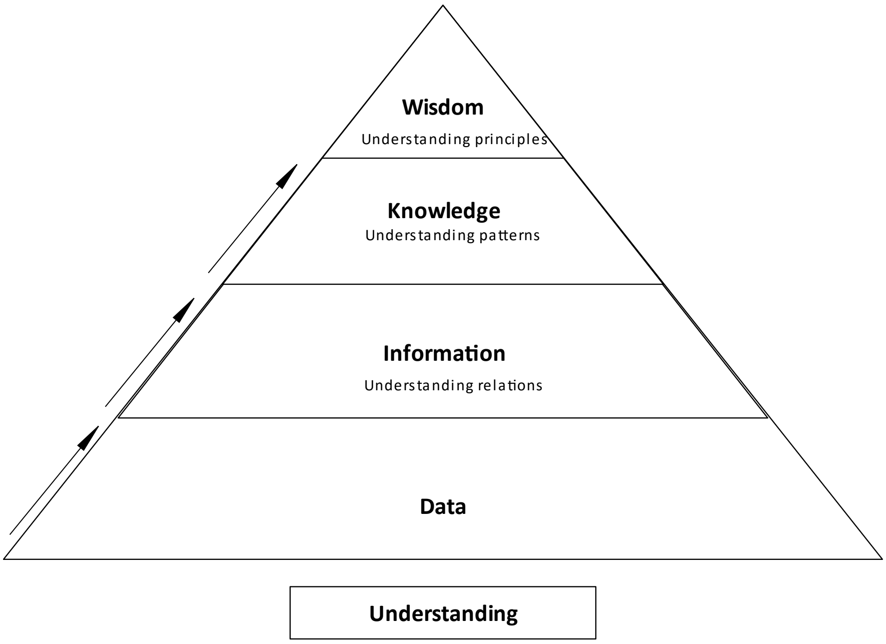

**Figure 2** Ackoff Data-Information-Knowledge-Wisdom hierarchy (Voronova, 2015)

### 2.2.3 Big data technologies & techniques

Communication service providers' main business is enabling data to be transferred between individual people, companies and the public sector. This has been possible through the advancements in communication technology made in over several decades. As their core competence is handling data in a trustworthy way, it can be assumed that CSPs are also in the optimal position to extract value from that data. Though without the right technology the task of utilizing that data can be overwhelming even for CSPs. One of the biggest challenges the CSPs may be facing is the integration of big data technologies to their existing IT infrastructure.

Hu et al. (2014) introduced a concept of big data value chain, which consisted of four phases; data generation, data acquisition, data storage and data analysis. The generated data is commonly comprised of unstructured data and requires real time analytics. Enabling this requires system architectures which support data gathering, storage, transmission and large-scale data processing mechanisms. Technologies like Big Table, Cassandra, Google File System, Hadoop, Hbase and MapReduce are used to aggregate, manipulate, manage and analyze big data (Wielki, 2013).

Apache Hadoop is an open-source software framework which supports large data storage and processing. It has been noticed by the industry and scholars alike to be one of the most substantial technologies in the big data movement. Hadoop can distribute the processing of large amounts of data on large clusters of commodity servers. Its scalability, cost efficiency, flexibility and fault tolerance make it particularly suitable for the management and analysis of big data. Figure 2 depicts a hierarchical architecture of Hadoop core software library. It covers the main function of big data value chain including data import, data storage and data processing (Hu et al., 2014.)

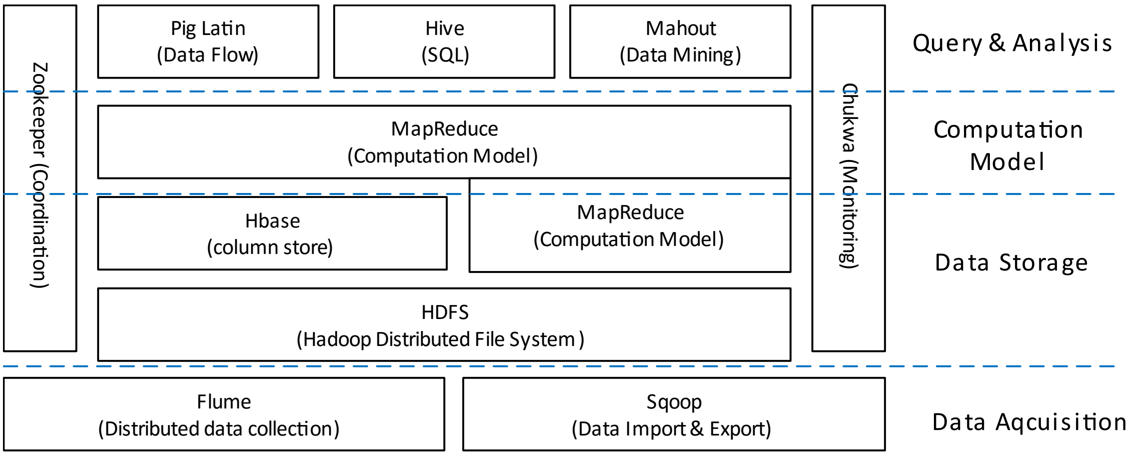

**Figure 3** A hierarchical architecture of Hadoop core software library (Hu et al., 2014)

In practice, reorganizing the data strategy and enabling big data technologies is a challenge for CSPs, but to fully embrace benefits of big data, the technology alone is not enough. CSPs must educate existing employees or recruit new talented data scientists or data analysts to manage the tools accordingly in order to extract the value out of the data (Wielki, 2013). Also, the management needs to be able to create needs, questions and problems which can be solved by the analysts who then would extract information out of the vast amount of data.

The implementation of big data technology enables different techniques to utilize and analyze data. These techniques can be e.g. data fusion and data integration, data mining, machine learning, predictive modeling, sentiment analysis, spatial analysis, simulation or time series analysis (Wielki, 2013). Presumably CSPs might be interested in especially in predictive modeling, since it could be utilized in predicting customer churn (Kamalraj & Malathi, 2013; Wei & Chiu, 2002; Manyika et al., 2011; Dam, 2013; Arora & Malik, 2015). Big data techniques also enable the collection of accurate and rich information for user profiling, especially due to their ability to process structured data as well as unstructured data in high volume from multiple sources (Hasan, Habegger, Brunie, Bennani & Damiani, 2013).

Implementation of new technology is not enough to utilize big data for a CSP, but merely a large part of the several challenges faced when becoming "big data compliant". CSPs need a holistic change not only in technological aspect, but also in business aspect as well. This would enable CSPs to benefit greatly from big data technologies as will be described on the next section.

## 2.3 Benefits of big data to the telecom industry

As aforementioned in the previous chapter, the ability to capture, store, manage and analyze large sets of customer generated data provides a significant opportunity for companies to extract valuable insight for supporting decision making and optimizing company business strategies. CSPs are in one of the best positions to utilize customer generated data, since their core business is maintaining and developing data networks. According to Dam (2013), CSPs are in the position to be able to capitalize on the insights locked inside big data and by managing to implement it swiftly, they can position themselves ahead of the competition, improve customer experience, drive new products, increase productivity, predict future trends, and create monetary value. CSPs who start to extract maximum value from their data assets can become fully two-sided businesses. Meaning that the CSPs can expand their revenue streams by many applications, among which are partnering with advertisers, retailers, health care and public administration (Dam, 2013).

Schroeder (2016) explored from an abstract viewpoint on ways how data can be a central component of a business model and the opportunities that it implies. Data business models are how enterprises can implement big data to their company strategy while being able to utilize several of the models. These data business models include *informing business decisions*, *data brokerage*, *data analytics as a service*, *consultancy and advisement*, and finally *provider of data related tools*. For a communication service provider the most likely business model to be implemented first would be of a data brokerage. Selling the first-party data to a third-party would enable profiting from gathering data even if the CSP lacks the capabilities to perform in their own data analytics.

According to Schroeder (2016) big data business models can be categorized in three different types; data users, data suppliers and data facilitators (Table 1). A CSP sits well with the data supplier type as a CSP could be able to gather data and even package it for sale. Being a data supplier might bring large efficiency gains, since the fixed costs of data production is usually high relative to the costs of distribution (Schroeder, 2016).

| Type | Example functions | Dependencies |
|---|---|---|
| **Data users** | Using data to inform strategic decisions; building data into products | Depend on suppliers for raw data, and on facilitators for infrastructure and skills |
| **Data suppliers** | Gathering primary data; aggregating and packaging data for sale | Depend on facilitators for infrastructure and skills, and on users both as customers and as sources of data |
| **Data facilitators** | Supplying infrastructure; consultancy; outsourced analysis | Depend on users and on suppliers as customers |

**Table 2** Big data business model typology (Schroeder, 2016)

As mentioned in the previous section, there are a collection of various big data techniques that can be used to discover knowledge in high volume, highly dynamic, and highly heterogeneous data. The big data techniques offer opportunities for user profiling that can result in very comprehensive user profiles (Hasan et al., 2013.). The process of user profiling consists of collecting information about a user to eventually construct their profile. The user profile may include information from numerous different attributes regarding the user, such as academic and professional background, geographical location, group membership, interests, preferences or opinions (Hasan et al., 2013). User profiles can be used for advantageous purposes like improving services, but also the data brokerage business model can be implemented by selling the profile information to third-parties.

The knowledge that a CSP might extract while gathering i.e. a user profile can be beneficial for the prediction of customer churn. The customers who leave their current telecommunication service provider and switch to a competing provider are referred to as "churners". According to Arora & Malik (2015) one of the earliest obtainable benefits, which a CSP can gain out of utilizing big data, would be the prevention of customer churn. Preventing churn can be done by predicting the customer's intentions when utilizing different kinds of prediction algorithms depending on the need (Kamalraj & Malathi, 2013). In the telecommunication industry, it is as much of an importance to convince new customers to subscribe to services as it is to prevent current customers from switching service providers in pursue of lower prices and/or better services. One way of reducing customer churn is to identify or predict potential churners and acting upon retaining them.

The reason for the desire of retaining existing customers arises from studies done on the subject. The studies state that acquiring new customers is expensive and it can cost up to five times as much when making a sale to a new customer rather than to an existing one (Kamalraj & Malathi, 2013). Understandably, acquiring new customers requires marketing and sales efforts which can be very costly to maintain for a CSP. Successful sales efforts are likely to be easier to perform towards customers who already are subscribers and are pleased with the service they get.

When a CSP acts as a data user and utilizes the gained insights to improve their own services, the subscribers might benefit as well. In general perspective, it has been suggested that by utilizing big data, organizations are able to create better and faster decision making and improve customer experience (Wielki, 2013). By proactively improving customer experience, a CSP can be able to reduce customer churn.

Most likely the first step for a CSP to benefit from big data is to utilize it while implementing the data brokerage data business model. This could provide early benefits while continuing a lengthy process of building big data compliant IT architecture and making necessary organizational changes and acquiring needed human talent to fully capitalize on the technological capabilities. Early benefits might tend to create a positive atmosphere also towards an organizational change that likely is needed while setting forward changes in IT infrastructure and strategic business models.

## 2.4 Challenges of Big Data in the telecom industry

The issues with implementing big data into the telecommunication industry are similar as seen across all industry verticals. Dhar & Mazumbar (2014) listed the main challenges that companies in all industries must likely face when adopting big data technologies. The challenges listed where the interoperability of IT-technologies, manageability of big clusters of hundreds of nodes, security of data, the maturity of big data technology providers, development scalability and maintainability of proper tools, and finally the reusability of the solution. The interoperability of IT-technologies is due to enterprises already having invested in business intelligence solutions and integrating big data solutions seamlessly to existing technologies is difficult and requires additional investments (Dhar & Mzumbar, 2014).

While implementing big data technologies the management of a large cluster of hundreds of nodes creates a problem for most companies, since a complete solution for managing and recovering big clusters is usually missing (Dhar & Mazumbar, 2014). As a company adopts big data technologies it also needs to adapt its organization to fit the new requirements.

Even if the implementation of the required technologies for big data analysis is successful, there is a shortage of analytical and managerial talent necessary to capture the full potential of big data (Manyika et al., 2011, Wielki, 2013). Making use of the gathered data requires specific skills which only well-trained data scientists can achieve. On top of the lack of appropriate skills, the organizations are likely to have internal structural barriers which create "organizational silos". Data connected to specific organizational functions (i.e. distribution, sales) are collected in "functional silos" rather than shared for the benefit of the entire company (Wielki, 2013). This type of challenge is likely due to company internal data sharing politics (Schroeder, 2016). It is important to ensure that the people who understand the business problems are able to use the right kind of data and be able to work with people who have the required problem-solving skills (Wielki, 2013).

Identifying the value in data requires a business mindset and a newer company culture which company managers might disregard due to old ways of working and habits. Therefore, the inability of senior management to view big data in a way in that it is strategically a key component of the company is noted to be a challenge as well (Wielki, 2013). The mindset of the managers should shift from making decisions on hunches and instinct to making decisions which are as data-driven as possible. Though data-driven decisions are not to be fully depended on, since data might not be providing the whole picture to a particular situation due to gaps in data signals or biases in data collection. Additionally, some managers might see that the effort of changing the company IT architecture to be big data compliant is too large compared to the benefits that it can create. This might be due to investments made towards software (some of which might be considered as legacy) already providing decent value by fulfilling current business requirements.

Even in the situation where the telecom provider is successful in gathering large amount of data in an efficient way, the complexity of data is a significant challenge. Today smartphones, personal computers, tablets, internet-of-things and various other sensors generate contextual and signaling data which might not have standardized formats. Schroeder (2016) describes that the big data users are facing challenges of poor data quality, missing knowledge of data context, lack of standards and accessibility towards the data and unstandardized regulations. Though the data might be plenty, there is an issue of the data quality being generally low.

Typically, data scientists might spend 75% to 80% of their time cleaning the data and preparing it for analysis (Schroeder, 2016). E.g. two datasets which are supposed to be merged might differ in their date format. Hence, the date format must be unified before merging the two datasets. Organizations expect to be able to analyze and react on information given from data in real time, but excessively long time taken in order to analyze huge data sets is a common problem (Wielki, 2013). Without the ability to react fast enough, the value of the data driven information deteriorates. If you were to receive tomorrows newspaper today it would be greatly valuable, but if you receive it tomorrow like everybody else it would contain information of insignificant value.

The most demanding challenges in big data arise from the legal aspect of handling different types of data. Some challenges relate to such issues as copyright, database rights, confidentiality, trademarks, contract law and competition law (Wielki, 2013). Therefore, CSPs need to remain transparent in their data collecting practices and ensure legality of data usage is maintained.

Most importantly, there is the issue of data security and privacy which is likely to be the most challenging and concerned problem in big data (Matturdi et al., 2014). The challenge pre-existed big data in the industry, but likely amplifies its significance due the new types of private insights which can be determined and gathered. The damage that a hacking attack creates might be increased significantly, since data would be gathered to data pools within the organization. The risk of leaking sensitive consumer data (i.e. consumer profiles, habits etc.) rises as data is combined from several sources. This aspect highlights the importance of maintaining sufficient company wide data security protocols and security software at the same time as the handling and analyzing of big data progresses. The following chapter will explore more on this issue.

# 3 PRIVACY IN BIG DATA

Personal privacy can be understood as everyone's own right to be left alone. It refers to personal information which has multiple dimensions: privacy of every individual's body, privacy of individual's personal behavior, privacy of personal communication, and privacy of personal data (Luo, 2002). This chapter will focus on the privacy of individual's personal behavior, communication and personal data as the privacy dimensions which might be affected by the big data techniques.

As mentioned in the introduction of this paper, O'Leary (2015) refers privacy as the ability of an individual to maintain or ensure information asymmetries over that individual's data and limit others from knowing or learning about them. Big data techniques are sometimes specifically used to gain insights about the behavior and attitudes of people when creating user profiles. Basically, the goals of big data and privacy seem to be fundamentally contrary to each other. As Shrivastva et al. (2014) noted, the most concerning challenge that big data needs to sort out is the privacy and security of information.

The former chapter explained how CSPs can utilize the data which they process for their economic benefit. In this chapter the privacy related issues are explored more deeply in order to get a more substantial look on what a CSP must keep in mind when seeking for those benefits.

## 3.1 Threats towards personal information

Shrivastva et al. (2014) define privacy as preventing the disclosure of sensitive information. Another definition of privacy is aimed at negative or "damaging" surveillance or unauthorized disclosure of a person (O'Leary, 2015). O'Leary (2015) explains that privacy refers to the freedom from damaging publicity, public scrutiny, secret surveillance, or unauthorized disclosure of someone's personal data or information, as by a government, corporation, or individual.

As mentioned before, probably the most challenging and concerned problem in big data is the issue of security and privacy (Matturdi et al., 2014).

The volume of information is a critical privacy parameter (O'Leary, 2015). As O'Leary (2015) explains, there is an uneven amount (volume) of data about some entities, agents and resources, and this indicates that more analysis can be done about some entities and agents compared to others. This results in more monitoring or surveillance being done on some individuals than others (O'Leary, 2015). E.g. the younger more "tech savvy" generation might subscribe to more digital services and generate more data than the older generation of people. Interestingly, Taddicken (2014), who performed a survey relating to privacy concerns on the social web, found the age of respondents having hardly any impact on the level of self-disclosure. Taddicken (2014) also concluded that users who are generally more willing to disclose much of personal information disclose the most on the social web. Then again, some people might value their privacy more than others and knowingly refuse to use some digital services and hence cognitively not generate as much data. Taddicken (2014) had a finding that people who prefer to stay self-contained actually reveal less information by resisting the invitation of the social web to disclose personal information. They might use add blocking tools and features enabling anonymity when e.g. browsing the internet.

Shrivastva et al. (2014) lists the major challenges of big data privacy being context-based privacy, co-related and aggregated data sets, threat modelling, privacy budgeting, and lastly policy and legal ramification. Whereas Matturdi et al. (2014) define one big data and privacy related issue as being the act of combining personal information of a person with external large data sets, which might lead to the inference of new facts about that person. There is a significant possibility that these kinds of facts about the person are secretive and the person might not want the data owner or any other person to know about them.

Furthermore, Terzi et al. (2015) describe in their survey that many big data projects are indicating the violation of people's privacy. According to Terzi et al. (2015) the increasing privacy concerns in big data include knowing new and secret facts about people by combining their personal information with other data sets. Organizations hence add value to themselves by collecting data from unaware people, threating illiterate people by predictive analysis of social media, tagging discriminated people by law enforcement, conflicting laws in different countries and lastly exchanging datasets between organizations (Matturdi et al.,2014; Terzi et al., 2015). This might sound as exaggeration, but at least companies have been known to exchange datasets between organizations and it is the business method when following the data distributor big data business model type.

Correspondingly, Jensen (2013) assumed two modes for the utilization of datasets of big data analytics, depending on the intention of its use. The first mode consisted on the verification of a pre-existing hypothesis and the second mode is that of identifying interesting facts hidden within the dataset. Two key concepts of aggregation of datasets within big data context were defined to be the aggregation of schematically identical datasets, and the linkage creation by joining datasets from disjointed contexts. Creating a linkage between different datasets require the use of some key information shared in both datasets which are data fields that have identical, similar, or otherwise sufficiently related values. Such values might be email addresses, postal codes, or combinations of IP addresses and timestamps. The usage of the identity of the service user for dataset linkage bears some challenges in the privacy of an individual (Jensen, 2013.). When others piece data together that normally is not together, loss of privacy can happen since doing so might provide insight about an individual.

Anonymization of the data has been discussed as the key for privacy compliant big data utilization. Yet, one of the major threats to privacy in big data analytics is the ability to perform "re-identification attacks" (Matturdi et al., 2014). The re-identification attack is an intentional act which refers to the ability of a huge dataset being explicitly scanned for correlations which lead to unique fingerprints of individuals. Linking different types of datasets together increases the uniqueness of each entry to the point that a link referring to an individual's identity can be established (Jensen, 2013.)

The re-identification attack has three sub-categories of attacks, which are correlation attacks, arbitrary identification attacks and targeted identification attacks. The correlation attack consists of linking a dataset of uniform data values to other sources to create more unique database entries. An effective correlation attack consists of linking additional datasets with at least one entry that is unique in its combination of data fields, or to the point that no two in the database have all data field values identical (despite the user identification). Correlating two datasets by userIDs creates more information per userID, allowing for a more precise analysis on the individuals behind the data (Jensen, 2013.).

A targeted identification attack is an attack where an adversary tries to find out more details for a given individual human being. It is only successful if it is possible to link some entries in the database to the identity at hand, with a sufficient level of probability. This type of an attack is the most threatening of the re-identification attacks, as they are likely to have the largest impact to an individual's privacy (Matturdi et al., 2014; Jensen, 2013).

All things considered, there is a need to develop big data techniques that can collect information for user profiles while at the same time respecting the privacy of the users (Hasan et al., 2013). Many aspects need to be considered to fulfill this need. The exchange of data with third parties should be done in a responsible manner where the CSP, which is selling the data, ensures that the data is used for its predefined and intended purpose. In addition, the data should be anonymized accordingly to requirements with no ability to re-identify the users. That is if the users wish to be unknown.

If a CSP would implement big data techniques to analyze and utilize the data generated by their service subscribers, the subscribers should be notified about it. In other words, the users should be made aware of the fact that their data is gathered and analyzed to bring out new insights of them or their behavior. Giving the subscribers a choice to either forbid or allow the usage of their data empowers them and more likely makes them more susceptible to the idea of giving a CSP control of their data.

## 3.2 Enhancements

It is possible that the existing telecommunication ecosystem will become the preferred and trusted keeper of personal data, since the industry has a good record in dealing with customers' data, the strict regulatory framework in place, and its inherent data protection obligations (Dam, 2013.) Privacy can be enhanced by the means of gathering only the required data for specific analyzing (O'Leary, 2015). With respect to the data volume being gathered, O'Leary (2015) explains that if there is less data, there is less information and potentially greater privacy.

Other privacy enhancing solutions have been proposed, e.g. the integrated Rule-Oriented Data System (iRODS), which addresses privacy challenges across a broad spectrum of communities, with differing institutional goals and security and privacy concerns, by providing the adopter community the ability to develop and deploy solutions for data management, which are specific to organizational needs (Matturdi et al., 2014).

Shrivastva et al. (2014) proposed that the differential privacy approach for big data privacy can fulfil most of the challenges that could be encountered while preserving privacy in big data. To put it simply, the differential privacy approach relies on the probability that the output of two datasets will be nearly the same. When nearly the same output is produced from two different datasets, the adversary is not able to determine the actual targeted data set by any quasi identifier. In contrast, Hasan et al. (2014) explained that the differential privacy approach can render some subsets of the randomized data less useful while preserving poorly the privacy of individuals.

Certainly, some of the enhancements in big data privacy do not stem from technical issues, but merely are based on legislation and organizational matters. The next section will focus on the regulations set for data owners in order to enhance the privacy of individuals.

## 3.3 Regulations

Regulations set by governments and other authoritative institutes create privacy and data protection laws to secure and ensure individuality of the members of its society. Jensen (2013) described three major challenges in creating privacy-compliant big data analysis. These challenges were interaction with individuals, getting consent, and revocation of consent and deletion of personal data.

According to Shrivastva et al. (2014), privacy can be preserved by imposing the rules and legality to an individual and an organization. In most big data contexts, it is necessary to collect and process information that is bound to specific individuals. Including personally identifiable information to the scope of big data analytics entitles each individual to be informed about every type of data processing that is involved in this data. That right, which for European countries is fixed in the European Data Protection Directive, must be granted on a per-request basis (Jensen, 2013.). A company, which gathers personal data for later processing, must disclose the data of an individual as well as provide details on the algorithms and processes that are involved in the big data analytics performed with them. Conversely, the legal obligation to disclose complex data mining algorithms might pose a challenge to big data processing, since they can be considered business secrets of the data analyst (Jensen, 2013.).

Regulations and privacy laws require the gathering of an informed consent from individuals prior to processing their data (Jensen, 2013). Oral and written pledges are the most common solutions to ensure security and privacy (Matturdi et al. 2014). This aspect has similar complexity issues as explained above. Providing an informed consent implies that an explanation of highly complex algorithms must be conveyed to each individual in an understandable way. Making everyone truly understand these techniques and algorithms is quite a challenge for the data analyzing organization (Jensen, 2013.) It is highly likely that the workings of complex algorithms and all intentions of data analyzing cannot be understood by all individuals, which enhances the responsibility factor of the CSP and the regulatory body.

According to European privacy laws like the General Data Privacy Regulation individuals have the right to decide to revoke their given consent for processing their personal data at a later stage and have the right to be forgotten by the data processor. The individual might decide to do so e.g. due to mistrust towards the data collector. The privacy law implies that all processing of personal data of the individual revoking consent must be stopped and the data must be deleted. The challenge in deleting individual personal data is in the possibility of that data been already spread widely among data collectors and analysts. Keeping track of a particular individual's data throughout big data analytics contexts can be met by organizational requirements e.g. the means of log files (Jensen, 2013).These requirements strengthen the need for big data systems to be privacy compliant by design, which likely slows down the implementation of big data technologies in CSPs.

The privacy laws are strict for a good reason, but they can be seen to decelerate the swiftness of big data implementation. Yet, in different perspective the CSPs have the possibility to stand out from the competition by being early in providing privacy compliant big data driven services to their subscribers as well as business partners. While traditional CSPs, with their existing and possibly outdated IT infrastructure, are coping to the new regulations, new emerging competitors in the industry might have the advantage of not being weighed down by legacy IT infrastructure. True privacy compliancy and big data driven services by design can be then achieved relatively swiftly.

To summarize, big data describes a data environment in which scalable architectures support the requirements of analytical and other applications which process, with high velocity, vast volume of data which may have a variety of different data formats and which may include high velocity data acquisition. The fundamental goal of big data is to discover insights from data which in turn can be a major privacy issue when the data subjects are individual people. In definition, privacy refers to the ability of an individual to maintain or ensure information asymmetries over that individual's data and limit others from knowing or learning about them, which is highly contrary to the goals of big data. Nevertheless, there is a medium where big data, when regulated and utilized responsibly, can bring extensive benefits for CSPs as well as for their customers. CSPs could implement data-driven business models while their customers could relish on the improved customer experience. To achieve the medium, CSPs must gain the consent of their customers for their data to be utilized in the manner that big data strives for. The historical position of CSPs, being a trusted handler private data (phone calls, messages etc.), may indeed foster the pursuit of CSPs towards big data compliant business models.

The next chapter will examine the empirical part of this thesis and present a research model relating to the perceived information risks that telecommunication service subscribers might experience.  The research model is tested by conducting a survey regarding on how service subscribers' consent to their data being analyzed and the results will be measured in respect of e.g. reliability, validity and significance.

# 4 EMPIRICAL RESEARCH

This chapter describes the methods in which the empirical study for this thesis was be performed. The research model and survey questions are based on a study by Libaque-Sáenz et al. (2016) examining the factors influencing risk perception of telecom customers in the Korean telecommunication market. The study as well as the literature review of this study imply that in general, there is a need for overall improvement concerning privacy-related issues in the telecommunication industry and that the service providers ought to raise the level of customer trust and improve their practices related to information privacy in order to raise customer satisfaction and decrease the concerns towards information privacy. In the light of these findings, the objective of this thesis' empirical research is to measure the survey results gathered in contrast to the original study and analyze them in the context of the Finnish market.

First the objective of the study is presented with the support of the research questions. After this the research approach and research method are explored by clarifying how the material was gathered, the content of the survey and analyzation of the survey material. Finally, the results are summarized and reflected upon in the final chapter.

## 4.1 Objective of the study

### 4.1.1 Research problem and questions

The objective of this study is to clarify the positive as well as the negative privacy related issues of the implementation of big data technology and techniques in the telecom industry. The subject of the privacy related issues are the private individuals who subscribe to telecommunication services. The following research questions where set for the examination of the research problem:

- What are the privacy related issues of implementing big data techniques in telecommunication companies?
- How do communication service subscribers consent to their data being analyzed?

The intend of this study is to highlight the aspects which should be taken into consideration by CSPs to produce firm and fair big data driven services while ensuring sufficient level of privacy. Next the research approach selected for this thesis is described.

### 4.1.2 Research approach

Quantitative research was the research approach selected for this thesis, because the intention of the study is to map the opinions of a large number of communication service subscribers towards the utilization of the data they generate. The quantitative research approach aims to give a general understanding of the subject at matter and drives to answer questions like how many, how much or how often (Vilkka, 2007, 14). Quantitative research can be argued to be a sufficient research method to get a broad spectrum of responses through different demographics of service subscribers and draw statistical conclusions about how they experience their privacy.

The quantitative research method is done in a form of an online survey in order to gather as much data as possible in a short period of time. Along with a method for gathering data, a quantitative research requires the analyzation of the data by applying a research model to verify the reliability and validity of the model and data.

## 4.2 Reliability and Validity

Reliability refers to the statistical measure of how the data from the survey instrument is reproducible. Meaning that the same results can be reproduced out of similar conditions every time the research instrument is used. Yet, it is expectable that gathered data from a survey will have some amount of error in it (Litwin, 1995.). To minimize the chance of a random error, the survey is aimed to gain as many answers as possible with the limited time to perform the study. The survey of this thesis was open for replies for 9 days and gathered a total amount of 176 answers. Typically, a quantitative research has a large number of respondents and it is recommended that the minimal amount would be at least 100 (Vilkka, 2007, 17). Additionally, to minimize the occurrence of measurement error, the survey will question the respondents' demographic information and hence add precision to the survey.

Validity of a survey aims to assess on how well it measures what it sets out to measure. The survey's goal is to measure how do the communication service subscribers consent to their data being analyzed. Content validity is a non-scientific measure of a survey instrument's accuracy. It is a subjective measure of how appropriate the survey items seem to an organized review (Litwin, 1995). Content validity of the survey may be enhanced by the supervisor of this thesis.

A challenging aspect of creating a survey is to make the respondent motivated enough to answer it with adequate thought and focus. Therefore, the survey had minimal amount of pretext in order to maintain the respondents' interest in continuing rather than losing interest due to heavy reading. The survey was promoted to take only 10 minutes to fill out entirely.

The short amount of pretext may have an effect to the validity of the research, since a respondent might not fully understand what information about them could be included in personal data. Yet, there has been a lot of discussion in the public media about personal information privacy during the last century and a basic knowledge of the content of personal information data should be known to the majority of the general public.

In addition to the lack of effort in educating the respondent for the survey, the respondent might feel impatient in replying to a survey and discard the importance of fully understanding the survey questions and their specific meaning. As personal information data might be comprehended in several ways depending on the respondents' viewpoint on the subject, some inconsistencies might be produced in the answers.

The survey questions are based on the hypotheses introduced on the next chapter. Each of the hypotheses are asked about in the survey using two to four questions to support their claim. All together 20 questions were used in the survey.

### 4.2.1 Hypotheses

In effort to investigate the second research problem several hypotheses are tested. The second research problem aims to answer how do communication service subscribers consent to their data being analyzed. To answer the problem, a research model was adopted from the study of Libaque-Sáenz et al. (2016). The hypotheses of the research model are justified by the literary review of prior academic articles. According to their research results, factors which affect the experienced privacy of telecommunication service subscribers can be numerous and these factors are tested on the research model. Libaque-Sáenz et al. (2016) validated their research model (Figure 4) to be able to test four dimensions of information practices, their interrelations and the effect of information practices on trust, information privacy concerns and perceived information risks.

The Figure 4 illustrates 7 different hypotheses and their theoretical negative or positive relations between factors. This also creates a path model which connects variables and constructs based on theory and logic. The research model defines the perceived information risks (PIR) to be the dependent variable due to an argument that risk beliefs can be affected by information privacy concerns, trust and information practices (perceived data control). The dependent variable is defined as individual's beliefs regarding potential loss associated with giving CSPs permission to use personal information. In turn, perceived policy awareness and perceived information protection are hypothesized to positively correlate with perceived data control. Next the hypotheses are described more in detail.

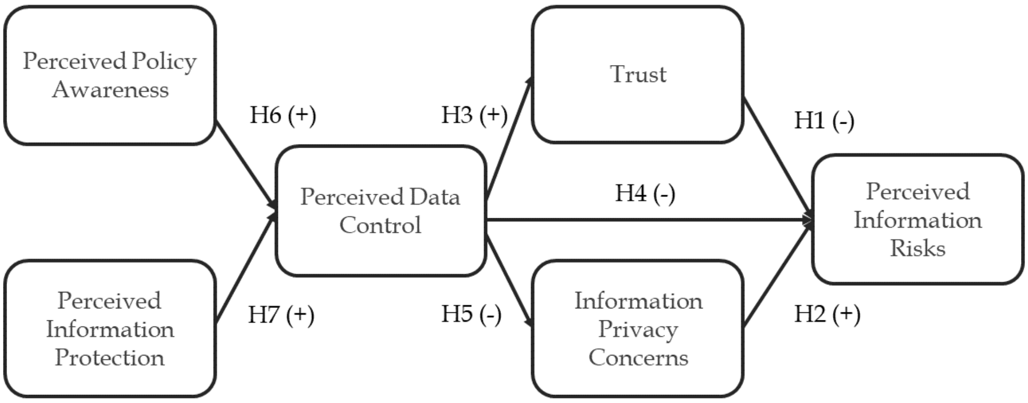
Perceived Policy Awareness
Perceived Information Protection
H6 (+)
H7 (+)
Perceived Data Control
H3 (+)
H4 (-)
H5 (-)
Trust
Information Privacy Concerns
H1 (-)
H2 (+)
Perceived Information Risks

**Figure 4** Research model (Libaque-Sáenz et. al, 2016)

First of the hypotheses aims to explain the correlation between Trust (TRU) and Perceived Information Risks. According to Pavlou and Fygenson (2006), trust is defined as individuals' beliefs that CSPs will act according to their expectations without exploiting their vulnerabilities. After CSPs have gained trust of their service subscribers, the consumers will develop lower level of risk perception towards the CSPs. Thus, the first hypotheses states that there is a direct negative relationship between trust and perceived information risks.

**H1**: Trust negatively influences perceived information risks

Secondly, the relationship between the information privacy concerns (IPC) and the perceived information risks is proposed by the second hypotheses.

**H2**: Information privacy concerns positively influences perceived information risks.

The privacy concerns variable defined here ques from Smith et al. (1996)'s description of an individuals' level of anxiety regarding CSPs information practices. As anxious people tend to perceive higher risks than non-anxious people, it is hypothesized that information privacy concerns have a direct positive effect on perceived information risks.

Perceived data control (PDC) is defined by Xu et al. (2011) as individuals' perceptions of their ability to control the usage of their personal information by CSPs. As people feel that they are in control, they are more willing to take risks and additionally they tend to estimate those risks as less serious (Brandimarte et al., 2013). Thus, the third hypotheses states that individuals' feeling of being in control positively influences trust towards the CSPs.

**H3**: Perceived data control positively influences trust

As individuals feel that they are in control of their personal data, the perceived measure of risks is lower, thereby having a negative relation towards the perceived information risks variable.

**H4**: Perceived data control negatively influences perceived information risks

Finally, the feeling of being in control of their own personal information also negatively influences information privacy concerns of individuals. Hence, the fifth hypotheses states that data control lowers information privacy concerns.

**H5**: Perceived data control negatively influences information privacy concerns.

The concept of data control is affected by the perceptions of information policy awareness and information protection. Perceived policy awareness (PPA) is defined as individuals' perceptions of the degree to which they understand CSPs information practices (Pavlou & Fygenson, 2006). For individual's, privacy policies are the assurances letting them believe that companies will behave accordingly (Xu et al., 2007). This leads to the sixth hypotheses to state that the policy awareness that individuals perceive makes them feel more in control.

**H6**: Perceived policy awareness positively influences perceived data control

Perceived information protection (PIP) is defined as individuals' perceptions that their CSPs have the ability to keep their personal information safe from security breaches (Pavlou & Fygenson, 2006). This means that when individuals believe that CSPs are competent to keep their personal information safe from security breaches, they will have higher perception toward their own ability to control the usage of personal information by the CSPs. Hence, the final hypotheses states that a higher perception of information protection leads to higher perception of data control.

**H7**: Perceived information protection positively influences perceived data control.

As for mentioned in the previous chapter each of these hypotheses are tested by two to four questions in the survey of this research. The next chapter describes the execution of the research more in detail by validating its content and listing the survey questions.

## 4.3 Research execution

As the empirical part for this thesis is done as a quantitative research, this chapter describes the collection of data via a survey, the content of the survey, questions (or measurement items) presented and ways the data was analyzed. The survey was created, published and initially analyzed on the online survey platform webropolsurveys.com. The survey was distributed by using an email list of students and faculty of the University of Jyväskylä, by word of mouth and posting a link to the survey on a messaging platform with a group of an around 100 participants. Additional interest towards answering the survey was sought by having a lottery for three gift certificates at the end of the survey.

Since the survey was sent to mostly a Finnish demographic, it had two language options (Finnish and English). The Finnish language version of the survey questions were added to this thesis as the second attachment. The next chapter lists the English language questions as the measurement items of the survey (table 2).

### 4.3.1 Survey content

Dinev (2014) emphasized that there ought to be more research in the anthropological and culture aspect of privacy since privacy patterns differ widely from one society to another. The survey of this thesis included demographical questions to enable the analyses of its findings to similar surveys done in different countries. Because the survey will be performed in Finland it is likely that the results will provide insight in the attitudes of Finnish people towards their data privacy.

The survey included the measurement items (table 3) presented in the study of Libaque-Sáenz et al. (2016). The questions were translated to Finnish language (appendix 2) in order for the respondents to have the ability to choose to answer the survey in either their native language or in English language. In the following table the measurement items are presented in the categories of their latent variable.

| | |
|---|---|
| **Information Privacy Concerns (CON)** | |
| **CON1**[a] | I am concerned that if I give permission to my Network Operator to use my personal information they have stored; this information could be misused |
| **CON2**[a] | I am concerned that others can find private information about me if I give permission to my Network Operator to use my personal information which they have stored |
| **CON3**[a] | I am concerned about giving permission to my Network Operator to use my personal information they have stored because of what others might do with it |
| **CON4**[a] | I am concerned about giving permission to my Network Operator to use my personal information they have stored because it could be used in a way I do not foresee |
| **Trust (TRU)** | |
| **TRU1**[b] | My Network Operator will be competent in using my personal information |
| **TRU2** [b] | My Network Operator will be honest in the ways it will use my personal information |
| **TRU3** [b] | My Network Operator will not seek to take advantage of me if I give permission to use my personal information |
| **Perceived Information Risks (PIR)** | |
| **PIR1**[a] | It would be risky to give permission to my Network Operator to use my personal information they have stored |
| **PIR2** [b] | If I give permission to my Network Operator to use my personal information they have stored, there would be high potential for privacy loss |
| **PIR3** [b] | If I give permission to my Network Operator to use my personal information they have stored, my personal information could be inappropriately used |
| **PIR4** [b] | If I give permission to my Network Operator to use my personal information they have stored, it would involve many unexpected problems |
| **Perceived Data Control (PDC)** | |
| **PDC1** [b] | If I give permission to my Network Operator to use my personal information they have stored, I will have control over who can get access to it |
| **PDC2** [b] | If I give permission to my Network Operator to use my personal information they have stored, I will have control over what personal information is used |
| **PDC3** [b] | If I give permission to my Network Operator to use my personal information they have stored, I will have control over how it is used |
| **PDC4** [b] | If I give permission to my Network Operator to use my personal information they have stored, I will have control over it |

| | |
|---|---|
| **Perceived Information Protection (PIP)** | |
| **PIP1**[a] | I expect my personal information to be adequately protected if I give permission to my Network Operator to use it |
| **PIP2**[a] | I feel secure that my personal information will be kept private if I give permission to my Network Operator to use it |
| **Policy Awareness (PPA)** | |
| **PPA1** [a] | I believe my Network Operator will disclose the way my personal information will be collected, processed, and used |
| **PPA2** [a] | I am confident that my Network Operator's privacy policy will have a clear and conspicuous disclosure |
| **PPA3**[a] | I am confident that I will be aware and knowledgeable about how my personal information will be used |
| **Scale:**<br>[a] **Strongly disagree / strongly agree**<br>[b] **Extremely unlikely / extremely likely** | |

**Table 3** Measurement Items (Libaque-Sáenz et al, 2016)

### 4.3.2 Analyzing the data

The analyzation of the gathered data was started off by initially skimming through the answers by using the reporting view of Webropol online survey software. Two of the responses were discarded from the material due to defective and improperly filled out survey forms. Only fully filled responses were approved for further analyses. All together 179 responses were submitted of which 2 of them being discarded due to insufficient filling of the survey. Next the data was exported from the online survey platform to an excel format. The excel format acted as the raw data to be used for further analyzation.

The analyzation of the raw data was initiated by using the SPSS 24 software to review the frequencies and descriptive statistics of the demographic. After the initial review of the demographical findings a confirmatory factor analysis was performed (CFA). The objective of the confirmatory factor analysis is to test whether the data is a good fit with the hypothesized measurement model (Kline, 2015 p.8). The structural equation modeling (SEM), which includes a diverse set of mathematical models (including CFA) (Ringle et. al, 2015), was performed using SmartPLS (v. 3.2.8) software. Partial least square approach for structural equation modeling (PLS-SEM) was preferred for this study despite covariance-based structural equation modeling (CBSEM) being considered the most appropriate statistical methodology when the goal is to further test and develop causal modeling where the theory is strong (Henseler et. al., 2009 p. 296). Additionally, it was noted that PLS-SEM has gained critique about its shortcomings as a method according to Rönkkö et. al. (2016).

Reliability of the construct was measured by coefficient alpha (Cronbach's alpha), which measures internal consistency reliability, the degree to which responses are consistent across the items within a measure (Kline, 2015 p. 69).

Since the coefficient alpha tends to provide a severe underestimation of the internal consistency reliability in latent variables in PLS models, it is appropriate to apply composite reliability measure, which takes into account that indicators have different loadings (Henseler et al., 2009 p. 299). Regarding the reliability of the measurement items, each indicator is assessed by their factor loadings.

Validity of the data is then examined by noting the constructs convergent validity and discriminant validity. Convergent validity is measured by as average variance extracted (AVE) as a criterion, which signifies that the set of indicators represents one and the same underlying construct and can be demonstrated through their unidimensionality (Henseler et al., 2009). Discriminant validity is measured by the Fornell-Larcker criterion. The criterion assumes that a latent variable shares more variance with its assigned indicators than with any other latent variable (Fornell & Larcker, 1981). The AVE of the latent variables should be greater than the latent variable's highest squared correlation with any other latent variable (Henseler et al., 2009).

The structural model was assessed by coefficient determination ($R^2$) of the endogenous latent variables. Additionally, the estimated values for path relationships in the structural model were evaluated in terms of significance. The analyzation described above will be put to context on the next chapter where the results for these aforementioned measurement instruments are presented.

# 5 FINDINGS AND CONCLUSIONS

This chapter describes the findings of the research survey and aims to explain them in the context of the theoretical framework. First, the demographic results were analyzed followed by examining the answers of the theoretical part of the survey. Subsequently, the survey results were investigated using the research methods related to quantitative studies. After coding the answer material, the correlations of the factors were analyzed and reflected to the general population.

## 5.1 Description of material

According to Vilkka (2007, 17) it is typical for a quantitative research that the amount of responses is large, but the recommended minimum amount for observations units is 100 if the research uses statistical methods. It is mentionable that the sample of the population never fully describes the population and due to that reason, the gathered results are only valid in some probability with the general population. Additionally, the sufficiency of the amount of responses depends on the methods used when analyzing the results. E.g. if the objective of the analysis is to compare different groups, the size of the sample should be no less than 30 from each group (Vilkka, 2007, 56-57.)

## 5.2 Respondents´ background

### 5.2.1 Demographical findings

The survey began by asking questions about the respondent's background. The demographical questions included inquiries regarding the respondents' gender, age, level of education and occupation. Additionally, the respondent was asked to select one or more service providers to which the respondent was subscribing to. The demographical questions disclosed the differences between the survey respondent sample and the general population of Finland.

Firstly, the survey gathered responses majorly from male respondents as 59,7% of the subjects were male and 40,3% female. This ratio does not comply with the population of Finland as the male gender only represents 49,3% of the Finnish population (Statistics Finland, 2018).

Secondly, age distribution between the respondents directed heavily to people in their mid-twenties. The youngest of the respondents were the age of 18 while the oldest were 57 giving a range of 39 years of age. The mean age was 29,43 and median was 27. Age was inquired by an open question. When comparing to the survey respondents' age structure to the age structure of Finland, provided by Statistics Finland (2018), it is evident that the survey findings point to the younger demographic rather than the whole of the Finnish population (table 2).

| Age | Survey respondents (%) | Age structure in Finland (%) |
|---|---|---|
| **15–19** | 2,8 % | 5,4 % |
| **20–24** | 25,6 % | 5,9 % |
| **25–29** | 36,4 % | 6,4 % |
| **30–34** | 15,9 % | 6,4 % |
| **35–39** | 6,3 % | 6,4 % |
| **40–44** | 8,0 % | 6,0 % |
| **45–49** | 1,7 % | 5,9 % |
| **50–54** | 2,8 % | 6,7 % |
| **55–59** | 0,6 % | 6,6 % |

**Table 4** Age structure of survey respondents & Finland compared (Statistics Finland, 2018)

The question about the level of education offered four prefixed options. Most of the respondents, 46,6 %, informed undergraduate degree as their highest level of education. Both upper secondary education and graduate degree options gathered about a fourth of the answers. One percent of the respondents had a doctoral degree.

The occupation was enquired by giving the respondent six options for answering. Students formed the vast majority among the respondents, as 60 % represented the student demographic when inquired about their occupation. Managerial employees (19%) and employees (16%) were the second largest occupational groups among the respondents.

| Variable | Category | Frequency | % |
|---|---|---|---|
| **Gender** | Male | 105 | 59,7 |
| | Female | 71 | 40,3 |
| **Age** | 15-19 | 5 | 2,8 |
| | 20-24 | 45 | 25,6 |
| | 25-29 | 64 | 36,4 |
| | 30-34 | 28 | 15,9 |
| | 35-39 | 11 | 6,3 |
| | 40-44 | 14 | 8,0 |
| | 45-49 | 3 | 1,7 |
| | 50-54 | 5 | 2,8 |
| | 55-59 | 1 | 0,6 |
| **Educational level** | Upper secondary education | 48 | 27,3 |
| | Undergraduate degree | 82 | 46,6 |
| | Graduate degree | 44 | 25,0 |
| | Doctoral degree | 2 | 1,1 |
| **Occupation** | Student | 106 | 60,2 |
| | Unemployed | 1 | 0,6 |
| | Employee | 28 | 15,9 |
| | Managerial employee | 34 | 19,3 |
| | Entrepreneur | 2 | 1,1 |
| | Other | 5 | 2,8 |

**Table 5** Respondents' demographic data

### 5.2.2 Service provider

Since the objective of the research was to investigate telecom customers' opinions on privacy, it was essential to enquire the service provider of the respondent in order to determine whether there were disparities in the views of different service providers' customers. The respondents were able to choose several options due to the possibility that an individual respondent might subscribe to more than one service provider simultaneously. Therefore, the amount of answers (242) exceeded the number of respondents (176). There are 4 major telecom companies in Finland competing for subscribers (Telia, Elisa, DNA and Moi). Saunalahti operates under the Elisa brand since it was acquired by Elisa in 2005. The two service providers were kept as separate answer options due to the possibility that a customer subscribing to Saunalahti's services might not perceive Elisa as their primary service provider.

The majority (59%) of the survey respondents subscribed to Telia while the second largest subscriber base was of Elisa (Elisa 36%, Saunalahti 16%). Also, two telecom providers (MTS in Russia and TNNet Oy) were mentioned in the responses both only once.

| **Variable** | | **Frequency** | **%** |
|---|---|---|---|
| **Service provider** | Telia | 103 | 58,5 |
| | Elisa | 64 | 36,4 |
| | DNA | 40 | 22,7 |
| | Saunalahti | 28 | 15,9 |
| | Moi | 5 | 2,8 |
| | Other | 2 | 1,1 |

**Table 6** Respondents' service provider data

Market shares of the Finnish communication service subscriptions (when writing this thesis) are divided relatively evenly between three major operators. According to the Finnish Communications Regulatory Authority (2018), Elisa is leading the market with 38% of the total number of subscriptions followed by Telia (34%) and DNA (27%). Rest of the operators have a total of 1% market share combined. Contrary to the survey responses Elisa had slightly larger market share in Finland though the majority of the survey respondents preferred Telia as their service provider.

## 5.3 Customer perception of privacy in the telecom industry

### 5.3.1 Measurement model reliability and validity

The reliability and validity of the model was measured by utilizing a measurement model. Reliability refers to the degree to which a set of indicators of a latent construct are internally consistent. CFA is used as a measurement tool to validate the measurement model. The measurement items where measured on a seven-point Likert scale. To begin with, the mean and standard deviation of the measurement items were calculated to get an overview on the general trends of the gathered responses and the scale of deviation within the items.

| Construct | Mean | Standard deviation |
|---|---|---|
| **CON** | 4,425 | 1,979 |
| **PDC** | 2,804 | 1,514 |
| **PIP** | 4,511 | 1,656 |
| **PIR** | 4,101 | 1,783 |
| **PPA** | 3,642 | 1,634 |
| **TRU** | 3,992 | 1,695 |

**Table 7** Internal construct consistency

The reliability of the survey subject factors was measured by examining their internal consistency with Cronbach's alpha and composite reliability. The internal consistency between the values of 0 and 1 portrays to which extent the elements in a test succeed in measuring the same concept and is thus affiliated to the inter-relatedness of the elements within the test. Generally, a value between 0,70 and 0,95 is perceived as acceptable, since a value below this may be caused by a low number of questions, poor relation between the questions or heterogeneous constructs (Tavakol & Dennick, 2011). In practice, Cronbach's alpha assumes factor loadings to be similar for all question items. Table 9 presents that the measurements for Cronbach's alpha generated results within the acceptable range of from 0,790 (Trust) to 0,946 (Information Privacy Concerns).

In turn, composite reliability does not assume factor loadings to be similar but reflects the varying factor loadings of the items. The more factor loadings fluctuate among the items, the higher the discrepancy between the values of composite reliability and Cronbach alpha. The classic reliability standard would be considered as a measurement of 0,70 or greater for a satisfactory composite reliability, yet acceptable reliabilities below this value maybe be obtained when an overall causal CFA model fit satisfactorily (Bagozzi & Yi, 2012). The data gathered for this study generated composite reliability values between 0,878 (Trust) to 0,961 (Information Privacy Concerns) as described on table 9. As aforementioned, standard satisfactory value for composite reliability can be considered as 0,70 or greater, which the measurements fulfill.

The convergent validity of the study, describing the degree of correlation between two measures within a concept, was researched by measuring the AVE (Average Variance Extracted) and factor loadings. According to Hair et al. (2014, 618-619), an AVE -value of ,50 or higher is considered to be adequate as it signifies that the factor explains at least half of the variation within the item. In this study the AVE -values fall in the range of 0,696 to 0,868 (table 9). The lowest AVE – values were received in the construct of Perceived Data Control (0,696) and Trust (0,706). This can be interpreted that even factors with the lowest AVE can still explain over half of the variation within them. If the measurement would have been below 0,5, it could have been interpreted that on average the constructs include more error in the items than variance explained by the latent factor structure set on the measure.

Discriminant validity is measured by the Fornell-Larcker criterion (table 8), which assumes that a latent variable shares more variance with its assigned indicators than with any other latent variable. According to Henseler et al. (2009), the AVE of the latent variables should be greater than the latent variable's highest squared correlation with any other latent variable. The bolded values on table 8 present the squared AVE's of variables, which are higher than compared to other latent variables.

| | IPC | PDC | PIP | PIR | PPA | TRU |
|---|---|---|---|---|---|---|
| **IPC** | **0,927** | | | | | |
| **PDC** | -0,291 | **0,834** | | | | |
| **PIP** | -0,479 | 0,388 | **0,932** | | | |
| **PIR** | 0,803 | -0,300 | -0,522 | **0,865** | | |
| **PPA** | -0,481 | 0,518 | 0,632 | -0,516 | **0,846** | |
| **TRU** | -0,463 | 0,565 | 0,635 | -0,481 | 0,662 | **0,840** |

**Table 8** Discriminant validity (Fornell-Larcker Criterion)

The coefficient of determination ($R^2$) is a measure of the model's predictive accuracy but can be also viewed as representing the exogenous variable's (i.e. Perceived Policy Awareness) combined effect on the endogenous variable(s) (i.e. Perceived Information Risks) (Hair et al.,2014). This effect ranges from 0 to 1 representing complete accuracy but rough estimates regarding acceptable $R^2$ levels or predictive accuracy are tolerably 0,75 (substantial), 0,50 (moderate) and 0,25 (weak). The $R^2$ measures of this study (table 8 & figure 5) range from IPC being the weakest (0,085) and PIR being the most substantial (0,659). The very low value of IPC implies almost insignificant predictive accuracy of the construct. The high statistical significance of the latent variables described by the t-values clearly exceed the minimum acceptable value of 1,96 (Hair et al., 2014 p.102).

| Construct | Cronbach's alpha | Composite Reliability | AVE | Construct items | t-value | $R^2$ |
|---|---|---|---|---|---|---|
| **Information privacy concerns** | 0,946 | 0,961 | 0,860 | CON1 | 80.936*** | 0.085 |
| | | | | CON2 | 67.579*** | |
| | | | | CON3 | 92.410*** | |
| | | | | CON4 | 61.025*** | |
| **Perceived data control** | 0,855 | 0,900 | 0,696 | PDC1 | 5.834*** | 0.274 |
| | | | | PDC2 | 34.933*** | |
| | | | | PDC3 | 24.448*** | |
| | | | | PDC4 | 42.475*** | |
| **Perceived information protection** | 0,848 | 0,929 | 0,868 | PIP1 | 50.019*** | - |
| | | | | PIP2 | 44.579*** | |
| **Perceived information risks** | 0,888 | 0,923 | 0,749 | PIR1 | 43.181*** | 0.659 |
| | | | | PIR2 | 65.027*** | |
| | | | | PIR3 | 20.133*** | |
| | | | | PIR4 | 29.116*** | |
| **Perceived policy awareness** | 0,798 | 0,883 | 0,716 | PPA1 | 42.152*** | - |
| | | | | PPA2 | 43.101*** | |
| | | | | PPA3 | 17.632*** | |
| **Trust** | 0,790 | 0,878 | 0,706 | TRU1 | 18.161*** | 0.320 |
| | | | | TRU2 | 64.549*** | |
| | | | | TRU3 | 20.066*** | |

*$p$<0.05, **$p$<0.01, ***$p$<0.001

**Table 9** Reliability & validity of the constructs

According to Henseler et al. (2009) the reliability of each measurement item should be assessed and that a latent variable should explain a substantial part (at least 50%) of each indicator's variance. The absolute correlations between a construct and each of its manifest variables should be higher than 0.7. In figure 5, all the measurement items correlate at loadings above the 0.7 threshold. Additionally, item loadings are larger than on any other constructs (Table 9).

| | IPC | PDC | PIP | PIR | PPA | TRU |
|---|---|---|---|---|---|---|
| **CON1** | 0,927411 | | | | | |
| **CON2** | 0,917793 | | | | | |
| **CON3** | 0,941484 | | | | | |
| **CON4** | 0,92245 | | | | | |
| **PDC1** | | 0,621701 | | | | |
| **PDC2** | | 0,907097 | | | | |
| **PDC3** | | 0,875916 | | | | |
| **PDC4** | | 0,89861 | | | | |
| **PIP1** | | | 0,935876 | | | |
| **PIP2** | | | 0,927285 | | | |
| **PIR1** | | | | 0,865819 | | |
| **PIR2** | | | | 0,909636 | | |
| **PIR3** | | | | 0,826449 | | |
| **PIR4** | | | | 0,857352 | | |
| **PPA1** | | | | | 0,889959 | |
| **PPA2** | | | | | 0,888555 | |
| **PPA3** | | | | | 0,753358 | |
| **TRU1** | | | | | | 0,806203 |
| **TRU2** | | | | | | 0,914237 |
| **TRU3** | | | | | | 0,795049 |

**Table 10** Confirmatory Factor Analysis

The hypothesized relationships between the constructs are measured by estimates provided by the path coefficients. The values are standardized on a range from -1 to 1. The coefficients closer to 1 represent strong positive relationships as the coefficients closer to -1 indicate strong negative relationships (Hair et al., 2014). Through this measure the most significant relationship in the model can be found in the positive relationship between information privacy concerns and perceived information risks with a value of 0,738 (Figure 5). This infers that the more there are information privacy concerns, the more a person perceives their being information risks. The other construct relationships did not indicate such strong relationships, with the weakest one being the relationship between perceived data control and perceived information risks. This relationship can be hardly considered significant, yet the relationship is negative as hypothesized in theory. By comparing the path coefficients to the hypotheses, it can be uncovered how the survey data supports the research model proposed by Libaque-Sáenz et al (2016).

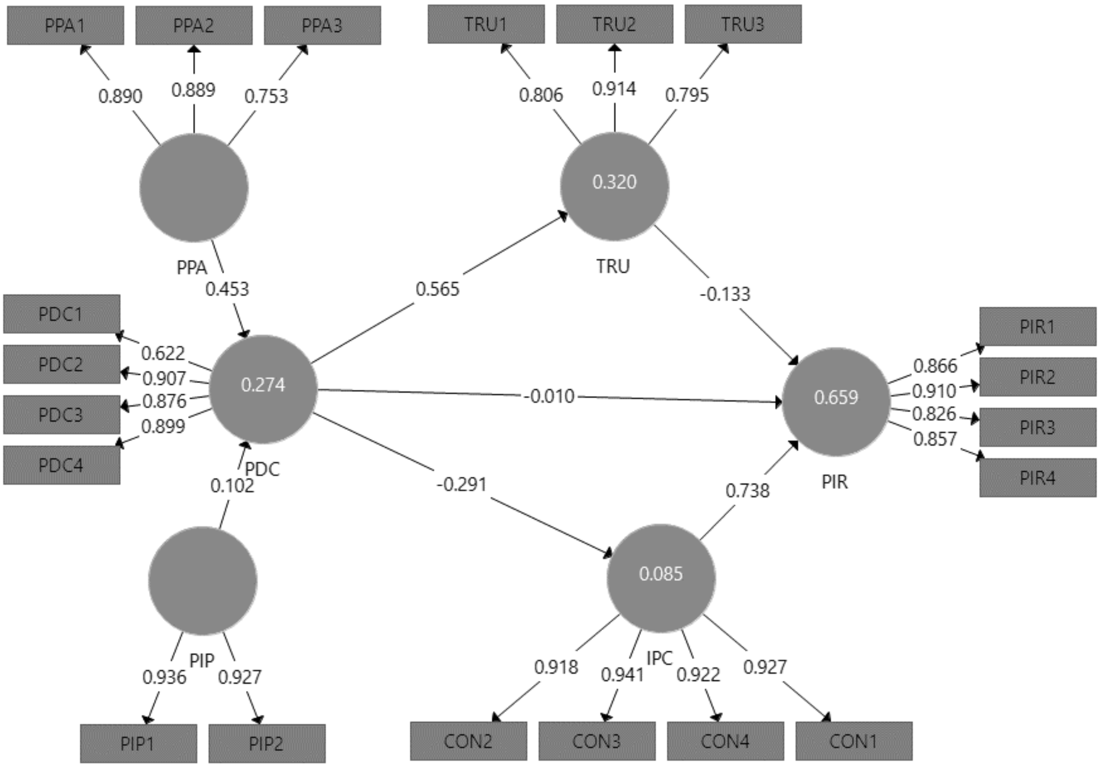

**Figure 5** Coefficients of determination ($R^2$), path coefficients & factor loadings

H1: Trust negatively influences perceived information risks

The relationship between the trust variable and perceived information risks is indeed negative, yet the significance should be examined as well according to Hair et al. (2014). The relationship significance is examined by t-statistics while a value greater than 1,96 (p<0,05), is considered somewhat significant and a value greater than 2,56 (p<0,01) then is very significant. The relationship significance between trust and perceived information risks is 1,978 (p=0,048) and is considered somewhat significant.

H2: Information privacy concerns positively influences perceived information risks.

The coefficient between information privacy concerns and perceived information risks is positively influenced and considered very significant with a 12,422 (p=0,000) significance measure.

H3: Perceived data control positively influences trust

The third hypothesis is again assuming correctly with a significance measure between perceived data control and trust at 10,450 (p=0,000) giving a very significant measure.

> H4: Perceived data control negatively influences perceived information risks.

The correlation between perceived data control and perceived information risks presented the lowest value out of all hypotheses. The significance of the relationship was merely 0,171 (p=0.864) which cannot be considered significant.

> H5: Perceived data control negatively influences information privacy concerns.

The fifth hypothetical relationship can be considered correctly assumed as the significance is 3,313 (p=0,001).

> H6: Perceived policy awareness positively influences perceived data control.

Perceived policy awareness does positively influence perceived data control, as the hypotheses states, with a significance indicating t-value of 5,921 (p=0,000).

> **H7**: Perceived information protection positively influences perceived data control.

The relationship between perceived information protection and perceived data control is not considered significant due to the low t-value of 1,093 (p=0,275).

In conclusion, most of the hypotheses were considered to be very significant (H2, H3, H5, H6), one was considered to be somewhat significant (H1) and two were considered not to be significant (H4, H7) meaning that the research model suggested by Libaque-Sáenz et al. does correctly describe the positive and negative correlations between latent variables, but two of the correlations can not be considered significant due to low performance in t-values. The results show that perceived information risks are determined by the individual's privacy concerns and trust, but contrary to what was hypothesized, not by perceived data control. Perceived data control has a negative correlation with perceived information risks, yet the correlation is not significant enough. The strongest influence on perceived information risks is from information privacy concerns, meaning that when an individual's concerns about information privacy grows; the perceived risks of the individual's information grows as well. Information privacy concerns and trust of an individual are determined by the perceived data control. When an individual's perceived data control is increased, the individuals experienced trust is very significantly affected in a positive way. Yet, when data control increases, the information privacy concerns of an individual decreases in a significant way. Finally, the perceived data control of an individual is positively affected by the perceived policy awareness, but the perceived information protection does not significantly affect perceived data control. Meaning, the more the individual perceives to be aware of the data policy in which data is handled, the more control of the data is perceived, but the protection of information the individual perceives does not increase perception of data control in a significant way.

In the following chapter the results of the measurement model and instruments are reflected upon. The conclusive chapter of this thesis considers all of the subjects mentioned previously in an attempt to answer the research questions introduced as the guidelines of this study.

# 6 SUMMARY AND CONCLUSION

Data privacy issues have been widely discussed in academia and in the general public during the last decade. In contrast, data driven business models have been a key interest of many industries which process customer generated data but are not fully utilizing the potential of that data. Big data technologies and techniques have been in the core of these discussions as the technology which enables future data driven business models. This paper has examined the privacy related issues of implementing big data techniques in telecommunication companies by exploring several academic publications. First the term "big data" was defined through number of different definitions found in the information systems literature. Then the emerging positive and negative issues regarding the implementation of big data in the telecommunication industry where introduced. Next, the meaning of privacy as well as the threats towards it, possible enhancements, and legislative regulations where examined. Finally, the empirical research of this thesis attempted to examine the factors influencing risk perception of telecom customers in the Finnish telecommunication market. An existing research model proposed by Libaque-Sáenz et al. (2016) was used and reflected upon through quantitative research methods. The progress of this thesis was guided by two research questions. The first research question was mainly answered in the literature review of this thesis while the second research question was answered by the empirical research.

The first research question of this paper sought to bring forward the privacy related issues of implementing big data techniques in telecommunication companies. The issues included the assurance of information asymmetries so that combining personal information with large datasets would not bring out new facts or sensitive information about a person which he/she does not want to be known. Culnan & Armstrong (1999) investigated prior research and concluded in a hypothesis which explains that people are willing to disclose personal information in exchange for some economic or social benefit subject to the "privacy calculus", an assessment that their personal information will later on be used fairly and they will not suffer negative consequences. Other major issues consisted of re-identification attacks as well as policy and legal ramification of data usage. In general, the issues were concerned of preserving privacy and information security of individuals while utilizing big data, and the regulations affiliated. Hence, there is a need to develop big data techniques that collect information while at the same time respecting the privacy of subscribers in question.

The reason why CSPs should confront these issues head-on, is due to the prospect which according to the literature, communication service providers, which implement big data capabilities and analytics to their technology infrastructure, can experience great benefits by doing so. CSPs can enhance their customer experience and create other valuable features through gathering and analyzing large amounts data that their customers generate. Besides the benefits gained from implementing big data, there is the added responsibility of data privacy that the telecommunication industry needs to keep in mind. Handling data privacy is not a new feature of the telecommunication business, but a feature which has been the backbone of the industry since its beginning. Due to the telecommunication industry's good reputation in handling privacy of their customers, it could be expected that big data would be implemented ultimately in a more privacy-conscious manner by the telecommunications industry than by other industries.

The regulations concerning data management can be seen as adding challenges to an already challenging field of technology. Before big data could be fully utilized to its maximum potential it is likely that the regulatory side of big data management will experience some sort of a reform. Admittedly, reaching the maximum potential in big data might not be the most reasonable goal in terms of privacy though it may in the process further privacy efforts. Regulations in their current state require that complex algorithms used in data analysis should be explained to the service subscribers so that they are understood. It is unlikely that every layman service subscriber can be educated to understand complex algorithms in an efficient manner. Yet, regulations are put into effect for good reasons, and such an important issue as privacy should be subjected to strict regulations.

One of the crucial aspects that can be extracted from the literature in regard to CSPs handling subscriber generated data is the control of the data. It can be reasoned that as long as the subscribers have control and knowledge of the data that they generate, the regulations are met to a large extent and privacy can be seen as partly fulfilled. Correspondingly, the empirical research of this thesis validated that perceived data control of an individual over his/her data significantly influences the trust of that individual via a positive correlation. As O'Leary (2015) explained, privacy refers to the freedom from damaging publicity, public scrutiny, secret surveillance, or unauthorized disclosure of someone's personal data or information, as by a government, corporation, or individual. When a subscriber holds the knowledge and control of that data he/she generates, there is essentially no secret surveillance or unauthorized disclosure of his/her personal data. Providing control of data to the subscriber would mean that the gathered data can be read, modified and removed by the subscriber in a simple way. This kind of a possibility for the subscriber can be considered reasonable.

It might be considerably less effort for new emerging CSPs to become big data oriented in their business compared to older and traditional CSPs. A new company can build up their IT infrastructure with data utilization at mind from the start. Building solutions which follow the modern data handling regulations may be easier to implement for a new CSP than for an older CSP which has the burden of the need to modernize old solutions. In addition, the employees of a traditional CSP might experience difficulties in understanding the new ways to providing communication services. The new data driven business mind set might be challenging for traditional CSPs to embrace and result in not doing anything disruptive to the industry. This may result in new companies to emerge which provide communication services in new and disruptive ways.

Admittedly, there is a dilemma with the goals of privacy and the goals of big data usage for business purposes. Big data drives to analyze data in large amounts in order to find out patterns and learn more about customers and their behavior. The goal of preserving privacy drives not to disclose this kind of information about individuals. To explore how the CSPs could be able to find a balance between big data and privacy, a survey study was carried out as an empirical research about the factors influencing risk perception of telecom customers in the Finnish telecommunication market.

The research model proposed by Libaque-Sáenz et al. (2016) acted as the theoretical framework of the empirical research part which consisted of an online survey addressed to communication service subscribers of the communication service providers in the Finnish market. As a theoretical contribution this thesis strives to further validate the research model in respect of Finnish culture as a representative of European privacy habits and culture. The model could have been extended upon by theories about new technology approval of the consumer but was not done so. The survey of 25 questions was directed to individuals of all ages but ultimately ended up representing an age region mostly of 20 to 30-year olds. The survey data was gathered using an online survey platform (Webropol.jyu.fi) which managed to gather 176 accepted answers. The gathered data was then used as the basis for examining factor correlations' significance of the research model. The correlations were studied by using the PLS-SEM method together with SmartPLS (v. 3.2.8) -software.

The factors were observed to be reliable and the factor correlations corresponded with the positive and negative correlations hypothesized in the research model. However, two of the correlations between latent variables (PDC/PIR & PIP/PDC) did not produce acceptable measurements regarding correlation significance. This despite that the same correlations were measured as significant in the research of Libaque-Sáenz et al. (2016). The result on the correlation being insignificant might be due to differences in the Korean and Finnish culture and communication service market, or the survey questions not being translated from Finnish to English while keeping the core intention of the question intact. Also, the lower answer amounts and narrow respondent age demography might have contributed in the lack of significance in these two correlations. In any event, the research model largely was able to predict the nature of the correlations and provide support for hypothesized claims.

In conclusion, the big data implementation in the telecommunication service industry holds a great deal of benefits for developing CSPs business models more towards making use of data they process, store and secure. Pursuing these benefits should not be in the cost of consumer privacy, since it might turn customers away from service providers and create churn for CSPs. Additionally, not following regulations might cause legal problems which can damage the business in terms of sanctions and reputation. With this in mind, the empirical research shows that the perceived information risks of the consumers can be reduced by increasing trust through giving consumers control of their data and increasing their awareness of the data policies that the CSPs follow. Likewise, the information privacy concerns, which positively affect the perceived information risks, can be reduced giving more data control to the consumer. The perceived data control of the consumer can be influenced by aforementioned awareness of data policies and less significantly the perception of information protection. By minimizing the information risks the consumers perceive, following regulations set by local authority and succeeding in implementing big data technologies and techniques (together with needed internal talent), CSPs can be able to transform their business models towards more data driven while maintaining and possibly even increasing their customer experience and loyalty.

# REFERENCES

Arora, D., & Malik, P. (2015, March). Analytics: Key to go from generating big data to deriving business value. In Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on (pp. 446-452). IEEE.

Bagozzi, R. P., & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. Journal of the academy of marketing science, 40(1), 8-34.

Brandimarte, L., Acquisti, A., & Loewenstein, G. (2013). Misplaced confidences: Privacy and the control paradox. Social Psychological and Personality Science, 4(3), 340-347.

Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. Organization science, 10(1), 104-115.

Debatin, B., Lovejoy, J. P., Horn, A. K., & Hughes, B. N. (2009). Facebook and online privacy: Attitudes, behaviors, and unintended consequences. Journal of Computer-Mediated Communication, 15(1), 83-108.

Dhar, S., & Mazumdar, S. (2014, June). Challenges and best practices for enterprise adoption of big data technologies. In Technology Management Conference (ITMC), 2014 IEEE International (pp. 1-4). IEEE.

Dinev, T. (2014). Why would we care about privacy? European Journal of Information Systems (2014) 23, 97–102.

Dinev, T., Xu, H., Smith, J. H., & Hart, P. (2013). Information privacy and correlates: an empirical attempt to bridge and distinguish privacy-related concepts. European Journal of Information Systems, 22(3), 295-316.

Emmanuel, I., & Stanier, C. (2016, November). Defining big data. In Proceedings of the International Conference on Big Data and Advanced Wireless Technologies (p. 5). ACM.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897-904.

Finnish Communications Regulatory Authority (2018), Matkaviestinverkon liittymät [Web log post]. https://www.traficom.fi/fi/matkaviestinverkon-liittymat [cited: 24.5.2019].

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics.

Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). A primer on partial least squares structural equation modeling (PLS-SEM). Sage publications.

Hasan, O., Habegger, B., Brunie, L., Bennani, N., & Damiani, E. (2013, June). A discussion of privacy challenges in user profiling with big data techniques:

The EEXCESS use case. In Big Data (BigData Congress), 2013 IEEE International Congress on (pp. 25-30). IEEE.

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In New challenges to international marketing (pp. 277-319). Emerald Group Publishing Limited.

Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. IEEE access, 2, 652-687.

Jensen, M. (2013, June). Challenges of privacy protection in big data analytics. In Big Data (BigData Congress), 2013 IEEE International Congress on (pp. 235-238). IEEE.

Kamalraj, N., & Malathi, A. (2013). A survey on churn prediction techniques in communication sector. International Journal of Computer Applications, 64(5).

Kline, R. B. (2015). Principles and practice of structural equation modeling. Guilford publications.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70), 1.

Libaque-Sáenz, C. F., Wong, S. F., Chang, Y., Ha, Y. W., & Park, M. C. (2016). Understanding antecedents to perceived information risks: an empirical study of the Korean telecommunications market. Information Development, 32(1), 91-106.

Litwin, M. S., & Fink, A. (1995). How to measure survey reliability and validity (Vol. 7). Sage.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Matturdi, B., Xianwei, Z., Shuai, L. I., & Fuhong, L. (2014). Big Data security and privacy: A review. China Communications, 11(14), 135-145.

O'Leary, D. E. (2013). BIG DATA', THE 'INTERNET OF THINGS'AND THE 'INTERNET OF SIGNS. Intelligent Systems in Accounting, Finance and Management, 20(1), 53-65.

O'Leary, D. E. (2015). Big data and privacy: Emerging issues. IEEE Intelligent Systems, 30(6), 92-96.

Pavlou, P. A., & Fygenson, M. (2006). Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior. MIS quarterly, 115-143.

Ringle, C. M., Wende, S., & Becker, J. M. (2015). SmartPLS 3. Boenningstedt: SmartPLS GmbH, http://www. smartpls. com.

Rönkkö, M., McIntosh, C. N., Antonakis, J., & Edwards, J. R. (2016). Partial least squares path modeling: Time for some serious second thoughts. Journal of Operations Management, 47, 9-27.

Schroeder, R. (2016). Big data business models: Challenges and opportunities. Cogent Social Sciences, 2(1), 1166924.

Shrivastva, K. M. P., Rizvi, M. A., & Singh, S. (2014, November). Big data privacy based on differential privacy a hope for big data. In

Computational Intelligence and Communication Networks (CICN), 2014 International Conference on (pp. 776-781). IEEE.

Smith, H. J., Milberg, S. J., & Burke, S. J. (1996). Information privacy: measuring individuals' concerns about organizational practices. MIS quarterly, 167-196.

Statistics Finland (Suomen virallinen tilasto, SVT): Väestörakenne [Web log post]. ISSN=1797-5379. Helsinki: Tilastokeskus [cited: 24.5.2019]. http://www.stat.fi/til/vaerak/index.html.

Taddicken, M. (2014). The 'privacy paradox'in the social web: The impact of privacy concerns, individual characteristics, and the perceived social relevance on different forms of self-disclosure. Journal of Computer-Mediated Communication, 19(2), 248-273.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International journal of medical education, 2, 53.

Terzi, D. S., Terzi, R., & Sagiroglu, S. (2015, December). A survey on security and privacy issues in big data. In Internet Technology and Secured Transactions (ICITST), 2015 10th International Conference for (pp. 202-207). IEEE.

Van Den Dam, R. (2013, October). Big data a sure thing for telecommunications: Telecom's future in big data. In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on (pp. 148-154). IEEE.

Vilkka, H. (2007). Tutki ja mittaa: määrällisen tutkimuksen perusteet.

Wei, C. P., & Chiu, I. T. (2002). Turning telecommunications call details to churn prediction: a data mining approach. Expert systems with applications, 23(2), 103-112.

Wielki, J. (2013, September). Implementation of the big data concept in organizations-possibilities, impediments and challenges. In Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on (pp. 985-989). IEEE.

Voronova, L., & Kazantsev, N. (2015, July). The ethics of big data: analytical survey. In Business Informatics (CBI), 2015 IEEE 17th Conference on (Vol. 2, pp. 57-63). IEEE.

Xu, H. (2007). The effects of self-construal and perceived control on privacy concerns. ICIS 2007 proceedings, 125.

Xu, H., Dinev, T., Smith, J., & Hart, P. (2011). Information privacy concerns: Linking individual perceptions with institutional privacy assurances. Journal of the Association for Information Systems, 12(12), 1.

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: privacy and data mining. IEEE Access, 2, 1149-1176.

Zikopoulos, P., & Eaton, C. (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.

# ATTACHEMENT 1

Respondents demographics

| | | **N= 176** | |
|---|---|---|---|
| **Demographic** | Category | Frequency | Percentage (%) |
| **Gender** | Male | 105 | 59,7 |
| | Female | 71 | 40,3 |
| **Age** | 15-19 | 5 | 2,8 |
| | 20-24 | 45 | 25,6 |
| | 25-29 | 64 | 36,4 |
| | 30-34 | 28 | 15,9 |
| | 35-39 | 11 | 6,3 |
| | 40-44 | 14 | 8,0 |
| | 45-49 | 3 | 1,7 |
| | 50-54 | 5 | 2,8 |
| | 55-59 | 1 | 0,6 |
| **Education Level** | Upper secondary education | 48 | 27,3 |
| | Undergraduate degree | 82 | 46,6 |
| | Graduate degree | 44 | 25,0 |
| | Doctoral degree | 2 | 1,1 |
| **Occupation** | Student | 106 | 60,2 |
| | Unemployed | 1 | 0,6 |
| | Employee | 28 | 15,9 |
| | Managerial employee | 34 | 19,3 |
| | Entrepreneur | 2 | 1,1 |
| | Other | 5 | 2,8 |
| **Telco** | Telia | 103 | 58,5 |
| | Elisa | 64 | 36,4 |
| | DNA | 40 | 22,7 |
| | Saunalahti | 28 | 15,9 |
| | Moi | 5 | 2,8 |
| | Other | 2 | 1,1 |

**Attachment 1** Respondents demographics

# ATTACHEMENT 2

Survey questions in Finnish language

| Factor | Survey question in finnish |
|---|---|
| **CON1** | Olen huolissani, että antamiani henkilökohtaisia tietoja voidaan käyttää väärin. |
| **CON2** | Olen huolissani, että muut tahot kuin operaattorini voivat löytää henkilökohtaisia tietojani. |
| **CON3** | Olen huolissani henkilökohtaisten tietojeni antamisesta operaattorilleni, koska en tiedä mihin muut tahot niitä käyttävät. |
| **CON4** | Olen huolissani henkilökohtaisten tietojeni antamisesta operaattorilleni, koska niitä voidaan käyttää itselleni tuntemattomilla tavoilla. |
| **TRU1** | Operaattorini on pätevä käyttämään henkilökohtaisia tietojani. |
| **TRU2** | Operaattorini on vilpitön käsitellessään henkilökohtaisia tietojani. |
| **TRU3** | Operaattorini ei käytä hyväkseen henkilökohtaisia tietojani. |
| **PIR1** | Olisi riskialtista antaa operaattorilleni lupa käyttää henkilökohtaisia tietojani. |
| **PIR2** | Jos annan operaattorilleni luvan käyttää henkilökohtaisia tietojani, on olemassa suuri mahdollisuus yksityisyyteni menetykseen. |
| **PIR3** | Jos annan operaattorilleni luvan käyttää henkilökohtaisia tietojani, niitä voidaan käyttää väärin. |
| **PIR4** | Jos annan operaattorilleni luvan käyttää henkilökohtaisia tietojani, se johtaisi moniin odottamattomiin ongelmiin. |
| **PDC1** | Jos annan operaattorilleni luvan käyttää henkilökohtaisia tietojani, voin hallita kuka pääsee niihin käsiksi. |
| **PDC2** | Jos annan operaattorilleni luvan käyttää henkilökohtaisia tietojani, voin hallita mitä tietoja käytetään. |
| **PDC3** | Jos annan operaattorilleni luvan käyttää henkilökohtaisia tietojani, voin hallita miten niitä käytetään. |
| **PDC4** | Jos annan operaattorilleni luvan käyttää henkilökohtaisia tietojani, voin hallita niitä. |
| **PIP1** | Oletan, että henkilökohtaisia tietojani suojataan riittävästi operaattorini toimesti. |
| **PIP2** | Uskon että henkilökohtaiset tietoni pysyvät salassa operaattorini toimesta. |
| **PPA1** | Uskon, että operaattorini on avoin tavoista joilla henkilökohtaisia tietojani kerätään, käsitellään ja käytetään. |
| **PPA2** | Luotan siihen, että operaattorini tietosuojakäytäntö on tuotu selkeästi ja avoimesti esiin. |
| **PPA3** | Olen täysin tietoinen siitä, miten henkilökohtaisia tietojani käytetään. |

**Attachment 2** Survey questions in Finnish