A Work Project, presented as part of the requirements for the Award of a Master Degree in Management from the Faculdade de Economia da Universidade Nova de Lisboa.

Big Data, Bigger Privacy Concern?

Nicolaus Dreyer 2999

A Project carried out on the Masters in Management Program, under the supervision of:

Luís Rodrigues

January 6, 2017

Abstract

In light of the rapid growth of big data applications in times where the *internet of things* is taking over personal privacy, this paper studies the area where data analytics and privacy concerns overlap. Identifying that anonymization and consent frequently do not suffice for user data, this paper also points out the weaknesses of regulations. A survey with 200 respondents showed that the awareness of big data capabilities caused significant privacy concern and willingness for (counter-) action, thus emphasizing that big data-driven firms should take a possible shift in user perception and behavior into account when formulating their strategy.

Keywords: *Big data, data analytics, information privacy, internet technologies*

Acknowledgements

The author would like to express his gratitude to all parties involved in the development and completion of this thesis, in particular his thesis advisor Luís Rodrigues, family and friends, and finally all participants of the survey, who may or may not have been falsely promised an invitation on his future yacht in the West Indies.

Table of Contents

1. Introduction

"Personal data is the new oil of the internet and the new currency of the digital world."

(Meglena Kuneva, 2009)

The quote of the former European Commissioner for Consumer Protection given in a keynote-speech seven years ago has never been more relevant than today. The developed world has now fully reached the digital age and firms and governments have embraced the collection and analysis of high volumes of data as a source of competitive advantage (Lund et al. 2013). The numbers speak for themselves:

The last years have seen an exponential increase in data generation. IBM estimates that "90% of the data in the world today has been created in the last two years alone" ("What is big data?" 2016). The market for data analytics is growing rapidly, too. Revenues of big data and business analytics are expected to grow by 50% from $122bn in 2015 to $187bn in 2019 (IDC 2016). The demand for data analytics is not unsubstantiated – according to a study by McKinsey, a retailer utilizing big data may have the "potential to increase its operating margin by more than 60 percent" (Manyika et al. 2011).

This rapid evolution of data collection has been enabled significantly by the proliferation of new technologies and internet connected devices (commonly described as the *internet of things* or *IoT)* as well as many offline services becoming digital, changing the way we communicate and share information. The tracks of data users leave across the IoT, or, rather, *digital foot prints*, are created both deliberately and without our knowledge, while also potentially containing personal information.

Needless to say, this has several implications for privacy; which include that (1) the collection of these vast amounts of data within the IoT, even if only accessible by certain few individuals, severely reduce the autonomy of the individual in regards to personal privacy decisions; (2)

through the advances of big data analytics, entirely new insights can be inferred from collected personal data with means of probability and huge reference datasets; (3) the growing complexity of software makes for an increased vulnerability in security for this data; this in combination with the rising number of connected devices results in a greater target area for unlawful data extraction.

The bulk of valuable data and the amount of opportunities to gather these have made it attractive for corporations and even governments to extract information with little regard for personal privacy, as seen in cases such as the highly controversial US surveillance program *Prism*, revealed by the whistleblower *Edward Snowden* in 2013 (Lam 2013). The resulting increase in privacy concern and threats for the individual therefore implies adverse effects for participating in the IoT and diminishes the overall user experience of many other services through the imminent or even transpiring violation of personal privacy.

The focus of the following thesis lies in the privacy concern resulting from corporations collecting and processing vast amounts of data generated and provided by individuals, who's concern, when significant enough, could cause a shift in user perception and behavior with negative and possibly even severe repercussions for these data-driven corporations.

1.1 Problem statement

Predicting, considering and confronting trends is part of the key challenges for strategic management and has been identified as a source of competitive advantage (Gluck, Kaufmann and Walleck 1980; Uphill 2016). In light of many once-assumed unshakable business giants whose ignorance towards the trend of 'going digital' had cost them significant market share in favor of disrupters that embraced this trend (Pisano 2014), it would be wise to take into account other possible future trends such as a consumer shift towards a more privacy-friendly environment as a response to aforementioned privacy concerns.

That privacy and its inherent security concern still holds true to this day could be noticed in, amongst others, the great protests on a global scale following the Snowden-disclosure, while the whistleblower himself was commended a hero (Cassidy 2013). Despite this violation of privacy occurring on a state-level, a misuse or blunder on corporate-level has likewise caused much unrest, with leaders of such firms finding themselves in a tight corner to offering a plausible explanation (such as *Steve Job's* rare apology in 2011, following much criticism against *Apple* after evidences emerged that *iPhones* and *iPads* had been saving location data of users (Helft 2011)). A challenge for big data-driven firms is therefore finding a balance between attending to this concern of individuals, who in many cases are also users of their service, and retaining their vital core competence of data analytics.

1.2 Objective of study

The objective of this study is to shed light on the upcoming challenges of firms using big data containing personal information of their customers or other individuals, thereby exploring the implications that big data capabilities have for personal privacy concerns. In this context, a key focus lies in the evaluation of the possible repercussions individual users' awareness and knowledge have for these firms. As a consequence, the following two research questions will be dealt with:

(1) "Does the concern of individual internet users in regards to their personal privacy gain significance with increasing awareness about the surge of data collection and the rise of big data capabilities?"

(2) "Should big data-driven firms be concerned about the possibility that users may set off a trend towards a more transparent and privacy-granting internet usage?"

A key objective is, hence, to assess the significance of the concern mentioned in (1), measurable through the scale of the subjective discontent about this matter, and from a certain point onward

also through the dimension of (counter-) action undertaken by users. Latter can be reflected, for instance, in observed changes in or even the total abandonment of customer usage of the firm's services. In the context of this study, the author differentiates data privacy concerns arising from security vulnerabilities (i.e. unlawful extraction of personal data through e.g. cyberattacks) from data privacy concerns arising from learning about individuals' features and attributes (profiling) through analytics (and ***not*** through unlawfully obtained data). The following elaboration will focus on the latter type.

1.3 Structure of study

As mentioned above, the following paper will attempt to provide an answer to the questions whether awareness about big data capabilities caused significant concern, and whether firms should subsequently anticipate a shift in user perception and behavior. Therefore, this paper will in a first step focus on providing a review of relevant research and related literature in order to unify existing knowledge in this field as well as to give a common basis for later discussion. For this, mainly peer reviewed, academic papers will be used, to ensure a certain quality of sources, while there will be a focus on papers published after 2010, due to the currency of the topic. The author begins by illustrating big data and its capabilities and discussing the difficulty of protecting and effectively regulating privacy matters. These topics will be essential to understand the privacy concern of individuals. As a next step, the author presents the methodology and the results of the empirical research on individual's internet privacy concerns. On the basis of the results, the author will identify key characteristics, of which the findings will be the basis of the subsequent discussion. Finally, a conclusion is drawn, with an attempt to answer the research questions formulated above and presenting the resulting implications for big data-driven firms.

2. Review of related literature

2.1 Capabilities and application of big data

Definition

The technology research firm *Gartner* offers a definition of big data as "high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation." ("What Is Big Data" 2013). In other words, big data involves the high-speed processing of vast amounts of seemingly unrelated data as a means to provide unprecedented insights for making decisions and/or achieve efficiencies. The characteristics that distinguish big data from other previously conventional data analytics are the '3Vs' (volume, velocity and variety), although some organizations and practitioners include 4 or even 5 Vs (including variability, i.e. an inconsistency of the data record; and veracity, i.e. a high variance in the quality of the collected data) ("Extracting Business Value From The 4 V's Of Big Data" 2014, "Big Data - What Is And Why It Matters" 2016).

Application on personal information

The diverse capabilities of big data cover the different levels that are also encompassed by most (but not all) other forms of data analytics:

- Descriptive analytics, providing a description of "what happened" ("Descriptive analytics" 2016)
- Diagnostic analytics; providing an explanation of "why it happened" ("Diagnostic analytics" 2016)

- Predictive analytics; providing a forecast or estimation of "what is going to happen" (Predictive analytics" 2016)
- Prescriptive analytics; providing a recommendation of "what should be done" ("Prescriptive analytics" 2016)

The diverse areas of application and its precision have lead firms of almost every industry to not only include big data analytics as a complement in a supporting business intelligence function, but to integrate it as a crucial component in their operations to enable data-driven decision-making, streamlining of supply chain and production, and targeted advertising of which the last always requires some form of user information.

In this regard, a controversial and frequently referenced example for more advanced targeted advertising capabilities and application of predictive analytics involved the case of a father, who after complaining about his daughter receiving coupons for maternity products from the retail chain *Target*, found out in this way that his daughter was pregnant (Duhigg 2012). What had happened is that Target, based on the purchase history of certain unique products, applied predictive analytics by the means of big data, which gave the daughter a significantly high pregnancy prediction score; thus knowing about her pregnancy before her own father did.

Ideally, big data capabilities should benefit both the corporation deploying big data analytics as well as the consumer. Such positive cases are, for example, observable with credit card companies who track customer behavior on a large scale and are thus able to notice suspicious activities and detect or even prevent fraudulences (White 2011).

Due to its potential of making inferences and predictions with incredible precision, big data has also been adopted by other proponents outside of the business context, from predicting natural disasters to winning elections (Kedmey 2015; Grassegger and Krogerus 2016). Following the

presidential election of 2016 the British data analytics firm *Cambridge Analytica* claimed a significant contribution at Donald Trump's victory (Grassegger and Krogerus 2016).

By applying methods from the area of psychometrics (a field of data-driven psychology), Cambridge Analytica combined it with various datasets (e.g. loyalty cards, information on memberships, media subscriptions and even ethnicity or religion) obtained legally against payment from a myriad of different sources. With this, Cambridge Analytica claims to have not only identified undecided voters with Republican tendencies, but also subsequently assigned them to profiles according to their psychological characteristics (having learned this as well in the data analysis process) in order to administer the most effective method for ad-targeting (Grassegger and Krogerus 2016).

Unfortunately for the concerned individual, inference of characteristics cannot be stopped simply with the withholding or the mindful treating of personal or attributable information. Various studies have found that communities of like-minded people form on social media – like in reality – so that, through the consent of a few to give information on their attributes, big data is able to infer the attributes of the other members of this community (Barocas and Nissenbaum 2014; Han et al. 2014; Mislove et al. 2010). In certain cases, the deliberate withholding of certain information paradoxically reveals certain other information (e.g. unease for a specific topic) and can be *almost* as effective to infer new information about the individual by comparing the behavior to a comparable set of people that did give their consent (Barocas and Nissenbaum 2014).

Big data analytic capabilities even go so far that datasets with anonymized personal information are not anymore secure from re-identification. (King and Forder 2016). With probability methods and running the anonymized profiles against other data sets with incomplete personal information, it is simply a question of seeking similar patterns and finding an incriminating similarity to match the anonymized profile to the individual again.

2.2 The challenge in limiting privacy concerns

Looking back at the capabilities and applications of big data illustrated shortly before, it is clear – even with our current comprehension of the privacy concept – that many of those activities involve intrusions into individual's personal privacy. In this sense, the *Information Systems Audit and Control Association* (ISACA) provides the following relevant approach to segment data privacy risks, or rather sources of privacy concern from big data analytics:

- privacy concerns arising from ***security vulnerabilities*** (or "*amplified technical impact*" (ISACA 2014, 10)),
- privacy concerns arising from ***learning about individuals' features and attributes*** upon collection and analysis of certain data,

  and lastly,
- privacy concerns arising from the ***capabilities of re-identification of individuals*** within aggregated, initially (semi-) anonymous data (ibid.).

Nevertheless, only the last two forms, which involve the action and capabilities of big data analytics, are relevant for the following sections, as data privacy concerns arising from security vulnerabilities lie beyond the demarcation of this study.

To understand more about individuals' privacy concerns however, it will not be sufficient to categorize privacy concerns by topics. *Mai* (2016) argues that in order to accommodate the mentioned advances and innovativeness of big data analytics' capabilities, there must be a shift in the way of thinking about privacy from a definition (i.e. "[the] characteristics of privacy" (ibid., 192)) towards a model (i.e. "how privacy works" (ibid., 192)). Previously, a violation of privacy simply involved the collection of data (monitoring as well as capturing) and could be defended by the prevention and control of information transfer and/or -outflow. These are covered in the two privacy models of *surveillance* and *capture* by *Agre* (1994). While the

surveillance model captures the contrast and (power) relation as well as "[the] tension between the watcher and the watched" (Mai 2016, 174), the capture model rather focusses on the act of "codification of activities" (ibid., 174) as the source of privacy concern.

However, neither are able to fully accommodate the implications of advances in predictive analytics. In the previous example of Target's pregnancy prediction, there was no privacy concern in the customer's consensual provision of purchase information. The issue laid in the analysis of this information to generate new knowledge about an individual with means of probability and inference. Mai (2016) therefore accommodates this by the means of a third model, the *datafication model*, in which the focus does not lie anymore on the observation and collection, but on the processing of data. This perspective allows to understand how privacy concerns may also arise from gaining new information that has never been in the possession of an individual, or that the individual could not have had the ability to withhold (see aforementioned example of inference of attributes in social media communities).

Regulatory issues

Fortunately for firms and unfortunate for the privacy of individuals, there are still only little regulations governing the processing of data. *Barocas* and *Nissenbaum* (2014) highlight that regulations have commonly focused on "two types of procedural mitigations [for privacy protection]: informed consent and anonymization" (Barocas and Nissenbaum 2014, 21). While informed consent involves the authorization of a person for another party to receive their information, anonymization does not. However, anonymization de-identifies this information by removing any personal identifiable information (e.g. social security number, full name, passport number etc.) so that it is no longer personally attributable to the individual or a person cannot be identified by this data. The problem in these two measures is that these two measures of mitigation of privacy concern have little to no effect when facing the re-identification and

inferring abilities of predictive analytics of big data, as certain information and identities can still be revealed (Barocas and Nissenbaum 2014).

Furthermore, Barocas and Nissenbaum (2014) go beyond pointing out the weaknesses of the current privacy protection measures to accommodate advanced data analytics and argue that a regulation relying on a 'notify and consent'-principle (as it is used in a number of countries) is not sufficient. For this, a comparison to patients in the field of biomedicine is drawn, where consent alone is not enough to apply certain treatment protocols. On the one hand, the reason lies in the patients' frequent lack of knowledge in this field; however, even if this was not the case, the so-called "transparency paradox" (Nissenbaum 2011, 36) applies, which implies that a textual simplicity cannot go hand in hand with a transparency of an explanation. Subsequently, the responsibility of making such a decision is (at least partially) delegated. It is therefore conceivable that - also for privacy concerns – governments hold the ethical duty of intervention in the interest of a greater welfare in the society by not letting the industry regulate itself and not allowing individuals to cast away their privacy through unenlightened consent.

However, this is not the only reason there is difficulty in developing and enforcing regulations for privacy concerns arising from big data. Another difficulty lies in enforcing data privacy laws across borders on the internet due to its international nature (Geller 2016). In addition to varying privacy laws of different countries, a further explanation – although greatly simplified – lies also in the simultaneously overlapping of jurisdictions. Like the differentiation between state laws and federal laws in the US can complicate matters, similar can be observed in the European Union (EU), where EU directives still have to be incorporated into each member country's national laws (Geller 2016). In some cases, countries have national laws that precede such given directives, as it was the case with the current *Data Protection Directive* (ibid., 22) Proposals of cooperation and developing regulations between the EU and the US have therefore proven to be difficult. While the EU strictly places the responsibility of regulation into the hands

of institutions, the US on the other hand has a rather industry-centered approach in privacy regulation, where the industry is left to self-regulation in considerable extents (MacDermott 2013; Mohamed 2016). An adverse result of the largely self-regulated industry in the US can be seen in the increasingly higher intrusiveness of data tracking (e.g. in the e-payment industry) (Mohamed 2016). A clash for cross-border regulation therefore arises due to the differences in the two approaches and "threatens the global interoperability of […] [perhaps even] the Internet itself" (MacDermott 2013, 8), implying similar difficulties in international regulation regarding privacy concerns from big data.

3. Methodology

Following the review of existing literature and research on the background of possible implications that big data capabilities have for personal privacy and privacy concerns, a quantitative research was conducted in the form of a survey, as a means to determine and assess the significance of user's concern regarding data privacy. Respondents were asked to give answers to a questionnaire stating their awareness and concerns of privacy issues arising from big data. The collected results are being presented in the following sections in a descriptive method, which will offer material for further analysis and discussion at the later point of this thesis.

3.1 Empirical research design

In the questionnaire respondents were asked to give 38 responses in regards to their internet and online usage as well as their awareness of big data capabilities and concern of data privacy issues. The respondents were given predefined responses to choose from, either by means of a 3- or 5-point Likert scale to indicate their position on certain issue (e.g. "***how*** important is it for

you that your political affiliation is kept private and not exploited by corporations?") or with a single response from multiple choices to express a specific range (e.g. "hours spent on the internet"). Responses on the Likert scale were preferred over other response types to allow aggregation as well as a more comprehensive quantitative comparison of the individual responses.

The questionnaire was split into several different topics: the first section deals with the internet behavior of the respondents, in which they were asked to estimate the time they spend online with internet connected devices, as well as estimate the frequency they use internet services such as social media platforms, online messengers, gaming and web browsing. These questions serve to suggest the degree of familiarity of internet services of the respondent, as well as to estimate the volume and variety of data they generate through their online behavior. The subsequent sections were aimed at learning the respondents' attitude towards various personal information and their awareness and concern towards privacy breaches including their likeliness of changing their online behavior in the light of such breach. These questions serve to assess the impact awareness and knowledge about data privacy issues had on their concern and how their concern impacted their subsequent behavior. This section therefore attempts to provide material to find an answer for the previously defined research question, which deal with whether awareness about big data capabilities caused significant concern, and whether firms should subsequently anticipate a shift in user perception and behavior.

To give insights on the diversity of the sample, respondents were asked to provide demographic information of themselves, i.e. gender, age (range), nationality and level of education. For the exact sequence, structure and questions of the survey, refer to appendix I.

3.2 Participants

The online questionnaire was distributed via social media as well as email in December of 2016 and responses were collected for a period of one week. At the end of the collection period a total of 200 participants had submitted fully completed questionnaires, so that a total of 200 questionnaires were eligible for further study. As the online survey tool did not accept the submission of incomplete questionnaires, the number of participants that had started but had not completed the questionnaire was not registered. Of the respondents, 35.5% were male and 64.5% were female. Regarding age, none were below the age of 18, 43.5% were between 18 – 24 years, 45.5% between 25 – 34 years, 2.5% between 35 – 44 years and finally, 8.5% were 45 years or above. Of the 29 different nationalities taking part in the survey, the strongest represented were German (38.5%), Austrian (17%) and Portuguese (9.5%). The distribution of highest obtained education level was as follows: 0% below high school, 7.5% high school, 50.5% Bachelors or undergraduate degree, 41.5% Masters or undergraduate degree and 0.5% with Ph.D. (or equivalent).

4. Analysis

4.1 Results

The results of the survey (see appendix II for all aggregated results) showed a moderate to high participation in the internet of things by the sample, which was important for the subsequent analysis. 71.5% of the respondents indicated they spent 3 or more hours online per day with connected devices. Also, an all-round usage of the individually presented online service categories (*online chats and instant messengers, social networking sites, gaming, web browsing, music and video, shopping and purchases, navigation and other location based*

services and *online banking/paying utility or other bills or taxes*) was given; on average over 90% used messengers, social media and web browsing on a daily basis. The least represented was online shopping, which respondents only engaged in less than once a week, on average.

In a further step, the awareness and knowledge of data analytics occurrences and capabilities was enquired. Many respondents showed significant awareness when they were asked to indicate (for each online activity separately), whether they knew their data is being saved and analyzed, and that new insights including undisclosed information about them is being inferred/concluded: Three-quarter were fully aware of this occurrence in social media, while an absolute majority (i.e. > 50%) was fully aware this occurred in online shopping, web browsing, navigation and location based services and music and video services. Higher unawareness for such activity was observed for the categories of gaming and banking/paying bills and taxes, where more than 20% of the respondents were fully unaware, compared to an average of approximately 10% full unawareness for other categories.

Respondents were asked to also give their awareness of the ability of data analytics is to infer attributes through the probability of likeliness of social media peers. Here, the overall awareness was very high, where only 6% claimed that they were fully unaware of this capability.

In regards to privacy concern, the survey showed varying importance for different personal attributes and information categories to be kept private and not being exploited by firms. Respondents were asked to give importance to following personal information: *preferred product and brand, hobbies and activities of interests, approximate income class, personal character traits and general mood/ well-being, political affiliation, physical health issues and condition, frequently visited locations,* and *relationship status and relationship quality*. Almost 46% indicated slight to no importance at all for their preferred brand and products and 40.5% for their personal hobbies and activities of interest. The most importance was attributed to information on frequently visited places as well as physical health condition and issues, where

more than 80% indicated *very important* to *extremely important*, notably more than what respondents on average indicated for personal character traits and relationship status and relationship quality.

The survey showed relatively clear results when respondents were asked to state the increase of concern for their personal privacy after having learned that the mentioned personal information could be inferred through data analytics. Only 9% claimed that their concern hadn't increased at all, while 38.5% even claimed their increase in concern was '*very much'* or '*extremely'*.

Respondents also assessed a high probability that this information would fall into wrong hands, where the average of the responses lies between '*moderately probable'* to *'substantially probable'*, while almost the absolute majority (49%) rated that regulations concerning the use and privacy of personal information to be somewhat insufficient.

Further, respondents were asked to indicate their likeliness of influencing others or changing behavior their behavior in light of their privacy concern. On average, responses lied between *moderately likely* and *very likely* (31.5% and 37%, respectively) when being asked how likely the respondent would warn and inform friends and family about their online behavior in regards to the privacy concerns of data analysis. Respondents were also asked to indicate how much more they would be more cautious or would change their online behavior for each online service separately after learning the capabilities of data analytics (*or how their behavior had changed, if they had learned this prior to the survey).* The most significant change in cautiousness or behavior was observed in the category of online banking and paying bills/taxes, where the average response was *moderately*, while for the categories of online messaging, social networking, browsing, shopping and navigation services the average lied between *slightly* and *moderately*.

Additionally, the results of the survey showed several trends: Naturally, respondents that had indicated that their concern had changed *very much* or *extremely* after finding out about the mentioned data analytic capabilities, generally also gave much higher importance of privacy for their personal information, compared to their counterparts who claimed that that their increase in concern was less (see appendices III and IV).

'Heavy users', or respondents that spent more than 5 hours with connected devices a day showed a higher awareness of data analytics capabilities than 'moderate users' (up to 5 hours a day) (see appendices V and VI). Also, these heavy users claimed to be more inclined to warn friends and family about the online behavior in regards to privacy concerns of data analysis (see appendices VII and VIII).

However, it was not possible to determine whether a specific frequency of engagement in a one of the internet activities was linked with a certain degree of awareness of data analytic capabilities (or vice versa). The reason lies in that there was either difficult comparability (e.g. almost all respondents were 'heavy users' of social media and online messaging), or that there were only very insignificant correlation.

4.2 Discussion

Reviewing the results, one can observe significant awareness of data analytics inferring personal information for most online activities, while this awareness did not fully diminish the use of the activity in most cases. One such example is the use of social media, which over 90% of respondents still use on a daily basis despite being well-aware of these practices. On the other hand, less awareness of such data analyses was given with the category of online banking and paying bills, where respondents indicated that the enlightenment of data analysis practices in this field would cause their activity to be far more cautious than in other online activities.

One argument that may support this matter is that the personal information involved in banking activity may be of higher value. However, the information of 'approximate income class' was rated as less important than information on 'relationship status and relationship quality', or 'personal character traits and general well-being', latter even being frequently shared on social media. Additionally, credit card bills and other transactions already reveal substantial information for the bank, customers of banks are generally well-aware about the profound knowledge banks have about them, so that a general mistrust of the bank in regards to personal information is unlikely. Therefore, the author argues that the use of data analytics may be misinterpreted by many that are not well-aware of these practices. The line between using data analysis for general operations and data analysis for dubious practices are more likely to become blurred when something of high value like one's fortune is at stake, particularly when individuals are not familiar with the topic. A further explanation could be that having the perception that data analytics are a common practice in a specific field can promote an acceptance and decreases skepticism towards it, as in the cases of social media platforms and online shopping – but not in banking.

What is essential for this paper's research is that the survey revealed that a fair amount of respondents had shown an increase in concern upon learning about what the capabilities of data analytics implied for personal privacy. Furthermore, the awareness had also triggered substantial cautiousness in online activity as well as the willingness to warn and inform others. The latter behavior may be assessed as extremely effective due to the value of opinions and appeal of friends and relatives. However, as the survey was not conducted with open questions, it was not possible to determine how such an *influencing* of families and friends would look like.

Also, the rating of privacy laws and regulation as (on average) somewhat insufficient by respondents has reaffirmed what has been revealed in our research on regulations. The

weaknesses in privacy laws may therefore become increasingly relevant when possibilities are being found on how to overcome cross-border issues and the current inability to regulate *how* data is processed (see datafication model; Mai 2016). In light of this, the author argues that a shift in user-perception may cause citizens to urge for the advancement of data privacy regulations, impacting big data-driven corporations even more.

It must however be noted that the responses in this survey should be taken with caution. When being asked about an opinion, respondents may have given responses about ***not*** what they personally feel, but what is expected from them. As an example, studies have concluded that respondents would be inclined to advocate for a stronger privacy protection in public surveys than what is actually wanted (Hong and Thong 2013; Smith et al., 1996).

5. Conclusion

The gathering of large volumes of seemingly unrelated data and feeding it in advanced algorithms has allowed big data analytics to generate new insights of unprecedented precision. It is not without a reason though, that its capabilities to learn about individuals to target and manipulate them, or its ability to even identify withheld personal information have caused controversies amongst proponents of privacy issues. However, the traditional methods of "consent and anonymity" for mitigation of privacy issues are not sufficient when confronted with big data. Furthermore, regulation of personal privacy issues in the digital age still has its weaknesses, given that there is no real regulation for how legally obtained data is processed. The questions this paper has set out to answer were whether the awareness about big data capabilities caused significant concern, and whether firms should subsequently anticipate a shift in user perception and behavior. The survey of this paper confirms the first question, revealing increases in concern upon awareness for big data capabilities. Further, it has been shown that a

fair increase of cautiousness in online activities occurred upon learning about capabilities of data analytics. Additionally, users indicated likeliness to warn and inform friends and family of privacy concerns arising from data analytics. Therefore, a shift of user perception and behavior is plausible and cannot be fully discredited.

Additionally, the author gives possible additional explanations for a differing levels of concern for personal information when data analytics is used in a specific field, which include the lack of knowledge about the topic, as well as the perception (*or the lack thereof*) that it is a common practice in that field.

Corporations pursuing or considering a big data-driven strategy should therefore opt for a transparent approach when dealing with personal information to avoid unnecessary privacy concerns. For instance, user sophistication (by educating individuals about the capabilities of data analytics) may support lowering skepticism towards big data practices, while clearly communicating the corporation's intentions will help build valuable trust.

Besides focusing on users, corporations may also anticipate advances in data privacy laws, that could eventually have found means to regulate *how* data is being processed with respect to privacy issues.

5.1 Limitations and outlook

Despite having a significant amount of respondents in this paper's quantitative research, an over-proportional number of respondents originated from Germany and Austria, fell into the age range between 18 – 24 and 25 – 34 and had a tertiary education as well. Therefore, the sample of individuals cannot be considered diverse enough to achieve a statistical significance to claim a global applicability of the results. Additionally, studies have concluded that respondents would be inclined to advocate for a stronger privacy protection in public surveys

than what is actually wanted (Hong and Thong 2013; Smith et al., 1996); therefore, in order to produce values that would accurately reflect the participant's assessment, the empirical approach would involve a complexity that surpasses the preset restriction for length and scope of this thesis. The author hence opted to simply present the data in its raw form to evaluate the results descriptively with room for further interpretation.

As a result, it would be conceivable to extend the question of ***if*** a shift in user perception and behavior is imminent for big data-driven organizations to a further research on ***how*** these organizations can adapt to this shift or even modify their business model. As mentioned in this paper, the concern about big data and privacy is also an ethical one. Consequently, an exploration of how governmental institutions – traditionally responsible for preserving social welfare – may tackle the challenges of implementing data and privacy regulations on a global scale as well as what guidelines must be developed to retain the privacy of users with a consideration on data processing.

This paper may therefore serve as a pointer for firms pursuing big data strategies to grasp the significance the issue of personal privacy concerns of users will have for them in the future, and may also be seen as an impetus for the necessity of further research in this field.

6. Bibliography

Agre, Philip E. 1994. "Surveillance And Capture: Two Models Of Privacy". *The Information Society* 10 (2): 101-127.

Barocas, Solon and Helen Nissenbaum. 2014. "Big Data's End Run Around Procedural Privacy Protections". *Communications Of The ACM* 57 (11): 31-33.

"Big Data - What Is And Why It Matters". .2016 *SAS*. Accessed December 30. http://www.sas.com/en_us/insights/big-data/what-is-big-data.html.

Cassidy, John. 2013. "Why Edward Snowden Is A Hero". *The New Yorker*. Accessed December 30. http://www.newyorker.com/rational-irrationality/why-edward-snowden-is-a-hero.

"Descriptive Analytics". 2016. *Gartner IT Glossary*. Accessed December 30. http://www.gartner.com/it-glossary/descriptive-analytics/.

"Diagnostic Analytics". 2016. *Gartner IT Glossary*. Accessed December 30. http://www.gartner.com/it-glossary/diagnostic-analytics/.

Duhigg, Charless. 2012. "How Companies Learn Your Secrets". *The New York Times*. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

"Extracting Business Value From The 4 V's Of Big Data". 2014. *IBM*. http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data.

Gluck, Frederick W., Stephen P. Kaufmann, and A. Steven Walleck. 1980. "Strategic Management For Competitive Advantage". *Harvard Business Review*. Accessed December 30. https://hbr.org/1980/07/strategic-management-for-competitive-advantage.

Geller, Tom. 2016. "In Privacy Law, It's The U.S. Vs. The World". *Communications Of The ACM*.59 (2): 21-23.

Grassegger, Hannes and Mikael Krogerus. 2016. "Ich Habe Nur Gezeigt, Dass Es Die Bombe Gibt". *Das Magazin*. https://www.dasmagazin.ch/2016/12/03/ich-habe-nur-gezeigt-dass-es-die-bombe-gibt.

Han, Xiao, Leye Wang, Noel Crespi, Soochang Park, and Ángel Cuevas. 2015. "Alike People, Alike Interests? Inferring Interest Similarity In Online Social Networks". *Decision Support Systems* 69: 92-106.

Helft, Miguel. 2011. "Jobs Says Apple Made Mistakes With Iphone Data". *The New York Times*. Accessed Decemeber 30. http://www.nytimes.com/2011/04/28/technology/28apple.html?_r=1.

Hong, Weiyin and James Y. L. Thong. 2013. "Internet Privacy Concerns: An Integrated Conceptualization And Four Empirical Studies". *MIS Quarterly* 37 (1): 275-298.

IDC,. 2016. *Worldwide Big Data And Business Analytics Revenues Forecast To Reach $187 Billion In 2019, According To IDC*. https://www.idc.com/getdoc.jsp?containerId=prUS41306516.

ISACA. 2014. *Generating Value from Big Data Analytics*. Rolling Meadows: ISACA. Accessed December 30. http://www.isaca.org/Knowledge-Center/Research/ResearchDeliverables/Pages/Generating-Value-From-Big-Data-Analytics.aspx.

Kedmey, Dan. 2015. "Here's How IBM Is Helping Towns Predict Disasters". *TIME*. Accessed Decemeber 30. http://time.com/3927087/ibm-weather/.

King, Nancy J. and Jay Forder. 2016. "Data Analytics And Consumer Profiling: Finding Appropriate Privacy Principles For Discovered Data". *Computer Law & Security Review* 32 (5): 696-714.

Kshetri, Nir. 2014. "Big Data′S Impact On Privacy, Security And Consumer Welfare". *Telecommunications Policy* 38 (11): 1134-1145.

Lam, Lana. 2013. "Edward Snowden: US Government Has Been Hacking Hong Kong And China For Years". *South China Morning Post*. Accessed December 30. http://www.scmp.com/news/hong-kong/article/1259508/edward-snowden-us-government-has-been-hacking-hong-kong-and-china.

Lane, Julia, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. 2014. *Privacy, Big Data, And The Public Good*. 1st ed. New York: Cambridge University Press.

Lund, Susan, James Manyika, Scott Nyquist, and Lenny Mendonca. 2013. "Game Changers: Five Opportunities For US Growth And Renewal". *Mckinsey & Company*. Accessed December 30. http://www.mckinsey.com/global-themes/americas/us-game-changers.

MacDermott, Siobhan and J. R. Smith. 2012. "The Future Of Privacy: A Consumer-Oriented Approach To Managing Personal Data Online". *Thunderbird International Business Review* 55 (1): 3-12.

Mai, Jens-Erik. 2016. "Big Data Privacy: The Datafication Of Personal Information". *The Information Society* 32 (3): 192-199.

Mai, Jens-Erik. 2016. "Three Models Of Privacy - New Perspectives On Informational Privacy". *Nordicom Review* 37: 171-175.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. *Big Data: The Next Frontier For Innovation*.

McKinsey. http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

Marr, Bernard. 2016. "21 Scary Things Big Data Knows About You". *Forbes*. Accessed December 30. http://www.forbes.com/sites/bernardmarr/2016/03/08/21-scary-things-big-data-knows-about-you/.

Marr, Bernard. 2015. "20 Mind-Boggling Facts Everyone Must Read". *Forbes*. Accessed December 30. http://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read.

Kuneva, Meglena. 2009. *Roundtable On Online Data Collection, Targeting And Profiling - Keynote Speech*. http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm.

Mislove, Alan, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. 2010. "You Are Who You Know: Inferring User Profiles In Online Social Networks". In *International Conference on Web Search and Web Data Mining*, 251 - 260. New York: WSDM 2010.

Mohamed, Nada. 2012. "The Do Not Track Me Online Laws: Creating A Ceiling When The Sky's The Limit And We Are Halfway To Heaven". *Information & Communications Technology Law* 21 (2): 147-154.

Nissenbaum, Helen. 2011. "A Contextual Approach To Privacy Online". *Daedalus* 140 (4): 32-48.

Pavolotsky, John. 2013. "Privacy In The Age Of Big Data". *The Business Lawyer* 69: 217 - 225.

Pisano, Gary P. 2014. "In Defense Of Routine Innovation". *Harvard Business Review*. Accessed December 30. https://hbr.org/2014/06/in-defense-of-routine-innovation.

"Predictive Analytics". 2016. *Gartner IT Glossary*. Accessed December 30. http://www.gartner.com/it-glossary/predictive-analytics/.

"Prescriptive Analytics". 2016. *Gartner IT Glossary*. Accessed December 30. http://www.gartner.com/it-glossary/prescriptive-analytics/.

Richards, Neil M. and Jonathan H. King. 2014. "Big Data Ethics". *Wake Forest Law Review* 49: 393 - 432.

Saeri, Alexander K., Claudette Ogilvie, Stephen T. La Macchia, Joanne R. Smith, and Winnifred R. Louis. 2014. "Predicting Facebook Users' Online Privacy Protection: Risk, Trust, Norm Focus Theory, And The Theory Of Planned Behavior". *The Journal Of Social Psychology* 154 (4): 352-369.

Uphill, Kevin. 2016. *Creating Competitive Advantage: How To Be Strategically Ahead In Changing Markets*. 1st ed. London: Kogan Page.

Smith, H. Jeff, Sandra J. Milberg, and Sandra J. Burke. 1996. "Information Privacy: Measuring Individuals' Concerns About Organizational Practices". *MIS Quarterly* 20 (2): 167.

Solove, Daniel J. 2013. "Privacy Self-Management And The Consent Dilemma". *Harvard Law Review* 126 (7): 1880-1903.

"What Is Big Data?". 2016. *IBM*. Accessed December 30. https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

"What Is Big Data? - Gartner IT Glossary - Big Data". 2013 *Gartner IT Glossary*. Accessed December 30. http://www.gartner.com/it-glossary/big-data/.

White, Martha. 2011. "Now Credit Card Companies Want Your DNA". *TIME*. Accessed December 30. http://business.time.com/2011/10/27/now-credit-card-companies-want-your-dna/#ixzz1ckBhcng8.

Young, Alyson Leigh and Anabel Quan-Haase. 2013. "Privacy Protection Strategies On Facebook". *Information, Communication & Society* 16 (4): 479-500.

Zhou, Wei and Selwyn Piramuthu. 2014. "Information Relevance Model Of Customized Privacy For IoT". *Journal Of Business Ethics* 131 (1): 19-30.