IADIS International Journal on WWW/Internet
Vol. 14, No. 1, pp. 70-85
ISSN: 1645-7641

BIG DATA PRIVACY CONCERNS IN THE LIGHT OF SURVEY RESULTS

Jędrzej Wieczorkowski . *Warsaw School of Economics, Al. Niepodległości 162, 02-554 Warsaw, Poland*

Ilona Pawełoszek, *Czestochowa University of Technology, Faculty of Management, Ul. Armii Krajowej 19 B, 42-201 Częstochowa, Poland*

ABSTRACT

Big Data may be understood as data sets whose sizes exceed the capacity of conventional database tools. The Big Data resources may include business transactions, e-mail messages, photos, surveillance videos and activity logs. Big data can be analyzed with the aim to draw informative results that lead to better decisions and strategic business moves. Although Big data could benefit many areas of social life and business, it also raises privacy concerns.

The paper discusses the issue of privacy and threats related to using big data technologies, especially personal data processing, video surveillance and monitoring the internet users' behavior during different activities. The aim of the paper is identification of subjective perception of privacy violation related to mass personal data processing. For this purpose the authors present the questionnaire survey results that was conducted recently among the students of Warsaw School of Economics.

KEYWORDS

Big data, data analysis, invasion of privacy, privacy concerns, surveillance, personal data

1. INTRODUCTION

The development of information technologies on one hand gives increased opportunities in business and the lives of individuals, but on the other hand it brings various threats. Both threats and opportunities are caused by new effective methods of mass data processing. Emerging capabilities for solving huge complex analytical tasks are known as big data phenomenon. The term itself has not been clearly defined yet but its analysis shows its multifaceted nature. The reviews of definitions of big data can be found in literature (Boyd and Crawford 2012), (Tabakow, Korczak and Franczyk 2014).

Big data may be understood as data sets whose sizes exceed the capacity of conventional database tools for gathering, storing, managing and analyzing data (McKinsey Global Institute 2011), or Big data is data that exceeds the processing capacity of conventional database systems, when the data is too big, moves too fast, or does not fit the strictures of database architectures (Dumbil 2012).

Some definitions underline the unstructured character of data, i.e. accordingly to Rouse (2011) big data is a general term used to describe the voluminous amount of unstructured and semi-structured data a company creates - data that would take too much time and cost too much money to load into a relational database for analysis. Other definitions refers to the type of processed data. According to PcMag encyclopedia (2016) Big Data refers to the massive amounts of data that collect over time that are difficult to analyze and handle using common database management tools. Big data includes business transactions, e-mail messages, photos, surveillance videos and activity logs (machine-generated data). Scientific data from sensors can reach mammoth proportions over time, and big data also includes unstructured text posted on the Web, such as blogs and social media.

The term big data is - driven to a large degree by the IT companies offering various types of solutions. However the problem is to define the term in a way that the definition would be timeless and set aside from the current state of technology. For example on blogs related to Microsoft (2013), an attempt to describe big data can be found as a term increasingly used to describe the process of applying serious computing power to seriously massive and often highly complex sets of information. According to SAS (2016) Institute big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it’s not the amount of data that’s important. It is what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves. In turn, Oracle (2013) in its report states the concept of big data refers to the number of basic groups of data, such as typical enterprise data (coming, for example, from ERP and CRM systems), data collected automatically (e.g. sensor data), data from the internet and social media.

In practice the 3V model of META Group (Laney 2001) is still considered as a basis of big data notion and its further development. In an original model coming from the report of META Group on the influence of electronic commerce, globalization and other trends on IT development, the three features were indicated which constituted the basis for the concept of big data that has been crystalized later, i.e.: volume, which means large amounts of data processed, velocity – which means variability of data, and variety – understood as heterogeneity of data. Recently many authors have attempted to indicate other characteristic features of big data, which could extend the “V” model, particularly: value – which means monetary worth of the processed data, veracity - credibility of the data, visualization – ability to visualize data.

The authors of hereby paper advocate the new possibilities of real-time (or near real-time) processing and processing ill-structured data are particularly important. The growth of the processed data is an evolutionary trend –technology development increases the number of data sets available. In contrast to the typical OLAP (On-Line Analytical Processing), in which the well-structured and aggregated data is used, the big data enables real-time analyses with detailed and often unstructured data sources. However, it is worth to notice that the data processing in data warehouses using ETL processes is often treated as big data (in such a case the Velocity feature is not present). Also the compound analytical processing of well-structured data is sometimes called big data processing (however the Variety feature is not present). These examples highlight the difficulties in distinguishing the big data phenomenon.

The phenomenon described above is associated with a threat of using new technology to the individuals' privacy violation. Threat to privacy is increasingly often perceived by IT users and other persons whose data is processed in information systems (Kamakshi 2014).

In this paper the problem is presented in the light of technology and business of big data. The aim of the paper is identification of subjective perception of privacy violation related to mass personal data processing. With this purpose in mind the survey was conducted which encompassed several questions potentially concerning privacy violation. Next the assessment of understanding the phenomenon and methods of big data was conducted. The paper also presents the literature analysis of the term big data and its three basic aspect: technological, business and social, with particular emphasis placed on the last one – related with privacy violation. The consecutive sections describe the research methodology and its results. Several statistical and data mining (classification trees) methods were used to describe and compare the survey results. The last section concludes the paper and presents most important findings, the propositions of further research efforts in this area were also indicated.

2. SOCIAL ASPECT IN THE 3-ASPECTS APPROACH TO BIG DATA

The authors of hereby paper identify three basic aspects of big data considerations (Wieczorkowski and Polak 2014):

- technological,
- business,
- social.

Distinguishing these three aspect has been the outcome of recent research conducted by the authors. The identification of the mentioned aspects was obtained through the research on a common understanding of big data term. The research encompassed, i.e. the analysis of the content of press articles on big data. The articles were sourced from non-scientific press and from various press releases.

The **technological aspect** is divided into two sub-aspects:

- information technology,
- analytical methods.

It represents a focus on the methods of big data analysis and information technology used. The big data concept is based on statistical methods, artificial intelligence, machine learning and data mining. This approach gives a possibility to analyze unstructured data such as texts from internet pages.

Big data requires specific solutions of very high computing performance and distributed processing. Acceleration of computing is associated with increasing computing performance (for example HPC - High-Performance Computing), extended memory use (in-memory), processing by database engine (in-database). The parallelization and distribution of computing is achieved by grid computing, cloud computing, applying MapReduce paradigm, in particular Apache Hadoop. The ineffective, old methods of data storage such as relational databases are replaced by new approaches. One of such solutions are column based NoSQL database management systems which are more convenient for storing weakly-structured data. Rapid growth of technological capabilities has become a starting point for development of big data in contemporary meaning.

The **business aspect** focuses on applications of big data. Shrinking cost of data gathering and processing makes it worth to process the data which were previously not possible or economically not feasible. The big data concept assumes almost costless data gathering and processing. Formerly, for instance, while designing analytical structures for data warehouses it was necessary to choose the most important business data in advance, and if there was too much data selected the cost and time of processing increased. The big data concept can be considered the next stage in evolution of business intelligence and predictive analysis.

Former applications of business intelligence were mainly associated with commercial sector. Undeniably big data for this sector creates wide range of opportunities, particularly according to Davenport and Dyché (2013) in: cost reduction, reducing the cycle time of analytical calculations, developing new product and service offerings based on data, supporting internal business decisions.

Applications of big data also seem to very promising in widely understood public administration and government. However until now the methods of processing large data volumes such as public statistics or public registers (population, ground, vehicles, companies etc.) have not been the typical big data in contemporary sense. The source of knowledge about public registers and other statistical data sources in Poland is the register of information systems for public administration (GUS 2013) which encompasses over 600 items. The trend can be observed for more real-time data while minimizing delays in processing, typical for public statistics. Most likely using such data sources will be increasingly close to big data not only regarding the data volume. Governments have increasing access to different data apart from typical public registers. Often it is weakly structured sensor data, such as city monitoring, more or less official internet traffic monitoring (Szymielewicz and Szumańska 2013). Moreover according to the reports of Tech America Foundation (2012) and McKinsey Global Institute (2011), and the Author's own observation it can be expected that big data analysis will proliferate to the areas such as: various frauds detection (in particular financial, economic, fiscal, or prohibited actions on securities market), public security (i.e. internet monitoring, wider use of monitoring systems), management of public services (in particular healthcare, transport, education and social care services), also in providing different information for supporting country government.

The **social aspect** concerns social repercussions of application of big data methods. Contrary to the aforementioned technological and business aspects, the social one goes beyond typical definitions of big data. The press research in popular journals indicate that the aspect of big data is very important in common public opinion. The subject has been touched often in popular press. The aspect of big data was raised most frequently in popular press. The social aspects of primary concern were the consequences of processing and using personal data, problems of infringements to privacy and threats of surveillance.

Currently the internet is an important source of personal data, where the social services are primary concern. The users' data may have significant value as they are potential customers for many companies. The possibilities of internet data analysis appeared long ago before big data was introduced, the benefits of internet data exploration were emphasized with the use of data warehouses (Pawełoszek-Korek 2008). Currently the basic business model of social services assumes making the social platform available to the community in exchange for access to the personalized information streams co-created and shared by the community (Polańska and Wassilew 2015). Advertisements, watched by the internet users can be based on the data profiled in the real-time. Banks and insurance companies may analyze community portals with the aim to profile the customers and to gain better assessment of their individual risk and creditworthiness. The mass personal data also come from mobile telephony services such as roaming, location data, logging to base station transceivers and Wi-Fi networks. Data on private financial transactions can be valuable, such as credit and debit cards payments, transactions from bank accounts, and data gathered by loyalty programs related to shopping.

Processing of very large data volumes and application of methods typical to big data can be also very useful for government, in particular for ensuring public security. Especially the role of police and emergency services can be important. Currently tracking internet requests is particularly important for prevention activities. For example an analysis of data coming from tax administration, customs shall and public registers, supported with internet tracking are applied for the detection of frauds. In case of prevention of serious threats the big value can be obtained from the resources of restricted access such as private messages and content of various files accessible in the cloud. Currently data of telecommunication operators is being extensively used to detect crimes. Other important sources of public security data are city monitoring, traffic monitoring, satellite and aerial photos. Monitoring provides sensor unstructured data that can be used with big data analysis methods. Apart from direct observation of people, particularly in places and -events particularly vulnerable to threats, the identification of vehicles is significant on the basis of their identification numbers.

The social aspect encompasses a **legal subaspect** regarding legal consent for using personal data, particularly with big data methods. The legal system is to ensure the minimal degree of protection of the right of privacy of the individual. Sensitive data concerning health, racial or ethnic origin, political opinions, religious or philosophical beliefs and sexual life are of great importance. The legal regulations of personal data processing address mainly businesses and administration (in particular state protection, public security and public finance management). In the scope of personal data processing for business purposes, legislation should ensure on one hand privacy protection, and on the other it should not hinder the usage of IT and economic development. In the area of state protection, the level of surveillance of individuals is an issue.

3. THEORY OF PRIVACY AND PREVIOUS RESEARCH

The subject of privacy is often raised in the context of comparison and a wordplay of two terms: Big Data and Big Brother. The latter refers to well-known novel by Orwell (Craig and Ludloff 2011), (Simon 2013). According to Craig and Ludloff (2011) privacy can be categorized into three basic types: physical (freedom of intrusion into one's physical person, possessions or space), informational (one's expectation of privacy when personal information is collected, stored, and shared in digital or some other format) and organizational (government agencies, organizations, and businesses expect to be able to keep activities or secrets from being revealed to others). From the point of view of this paper the second meaning is important. The significance of information privacy grows along with the advancements of IT, but it is not only related to big data phenomenon. Nevertheless the development of big data applications significantly influenced the perception of information privacy. Haire and Mayer-Schönberger (2014) note that currently core strategies to insure privacy (such as: individual notice and consent, opting out, anonymization) have lost much of their effectiveness in the light of the possibilities brought by big data. For instance contemporary computing power to a high degree facilitates the process of anonymization. Therefore it can be assumed that significant change in the approach to privacy is taking place. The question remains, how far the change in technology influences changes in the awareness of the problem.

In the literature a distinction is drawn between the notions of “privacy” and “right to privacy”. The first one describes what privacy entails and how it is to be valued, while the latter refers to the extent to which privacy is and should be legally protected (Solove and Schwartz 2009). Research on privacy must not be limited solely to legal aspects, because law may be imperfect or not facing up the reality.

Historically the first mature concept of privacy is the theory described as the “right to be let alone” (Warren and Brandeis 1890). Then, the attention was paid to the importance advances in communication methods and photography, which could affect someone’s privacy. The distinguishing between private and public life is important. Disclosing private life is inadmissible when it is not related to public interest. Later the notion of privacy was gradually expanded and various types were distinguished. For example (Solve 2002) argues that the many conceptions of privacy can be divided into six general headings:

- the right to be let alone;
- limited access to the self, or the ability to shield oneself from unwanted access by others;
- secrecy, or the concealment of certain matters from others;
- control over personal information, or the ability to exercise control over information about oneself;
- personhood, or the protection of one’s personality, individuality, and dignity;
- intimacy, which is to say, control over, or limited access to, one’s intimate relationships or aspects of life

Some of the above categories become particularly important in the light of modern information technology.

Nissenbaum (2009) divides the concerns over new technologies into three categories:

- monitoring and tracking,
- dissemination and publication,
- aggregation and analysis.

Actually all the above categories are to some degree related to the exploitation of big data. For example „monitoring and tracking” are used for capturing data for further processing. „Dissemination and publication” encompass the problem of facilitating the access to data and gathering historical data which are also the basis for further analyses. Particular attention should be paid to the category of „aggregation and analysis” to which the endurance of mass data should be included, as well for acquiring aggregated data as analysis of individual data (for example to create users’ profiles).

From the point of view of this work, privacy can be understood particularly as a control of private information flow. In such way the privacy is represented by the concept of Nisseubaum (2004), who underlines the problem of control to access to one’s own private information in some social context – which information, to whom, when and in what situation can be available.

The problem of attitude to privacy has been addressed in many researches since the times when completely different methods of data processing were used. An overview of types of the types of attitudes to privacy was presented by Kołodziejczyk (2014), based on previous empirical researches of different authors. Three attitudes to privacy are typically distinguished in society, these are: careless, pragmatists and fundamentalists. Regarding the method of research, the proportions between these groups are variable.

Westin (1996), when the capabilities of IT were significantly different than now, surveying the consumers market, identified about a half of respondents to the group of pragmatists – prone to share their personal data calculating risk and advantages. The remaining two groups, both consisting approximately 25% are fundamentalists who share their data reluctantly, and unconcerned **-** the group of careless who do not care for their personal data.

Sheehan (2002) basing on the research of Smith (1992) notes to the fact that after in-depth interviews, the persons who have limited knowledge of privacy begin to realize the problem and exacerbate their approach and moving, for example, from being pragmatics to fundamentalists. Also in this research the pragmatics are definitely the largest group but at the same time quite differentiated in details of their attitudes. Results of researches also show that the group which is most differentiated internally are young people (18-24 years old), which more often represent extreme attitudes. The level of education, in turn, influences increased criticism of data disclosure. Generally, the mentioned researches show that at least half of the society are pragmatists, prone (although not without criticism) to share their data having in mind advantages and context in which it is used. The problem of different attitudes to privacy was also stressed by Marwick et al (2010).

Currently quite popular subject of research related to privacy is threat to privacy on the Internet, particularly in social networking websites and their privacy settings. The privacy in social networks and usage of privacy settings has been discussed by Surma (2013), the author indicates that active Facebook users who understand the issue, also willingly use available privacy settings. The question arises what is the relationship of the users' knowledge of contemporary mass data processing technologies and feeling threat to one's privacy resulting from using these technologies. The research of Kołodziejczyk (2014), reveal that even students very often do not know how to use privacy settings in social networking portals. From the point of view of this paper, the two contexts of privacy in the Internet distinguished by Kołodziejczyk are particularly interesting, these are: social and institutional. The social context encompasses the threats related to individual recipient of information – usually family and relatives. Institutional context is related to institutions as potential recipients of private information, which can be public administration, advertisers etc. For most of the recipients more important is social context related for example to the access of known but unauthorized person to someone's private data. Probably the threats to privacy related to the institutional context seem to be more enigmatic and abstract, however they are discussed in this paper because the different institutions have the capabilities of processing big data. The research efforts can be found in literature concerning privacy issues related to big data methods (Victor, Lopez, Abawajy 2016). In our paper the authors try to deepen that problem.

4. RESEARCH DESIGN AND METHODOLOGY

Widespread interest in social aspect of big data induced the authors to undertake a survey concerning threat perception arising out of the violation of privacy. In contrast to previous research, the authors of this paper do not concentrate on general concept of privacy, we also do not limit to privacy on the Internet. The starting point of the presented research is data processing currently getting popular and called big data, in particular mass processing of personal data with the use of IT.

The method of in-depth direct interviews or questionnaire survey was considered, in which some actions related to the phenomenon of big data would be suggested to the respondents. The first method in similar research (for example conducted also among polish students, such as the research of Kołodziejczyk (2014) do not require in advance listing the actions which could threat the feeling of privacy. The lack of such suggestion can be considered disadvantage of this method because of the scope and multi-aspect character of big data. Therefore the authors decided to prepare the list in advance and in consequence the questionnaire survey with closed questions was chosen. The questionnaire begins with the part considering understanding of the big data phenomenon. From one hand this approach makes the respondents think over the subject of big data before answering the questions on privacy what can influence the results. On the other hand on the basis of the first part of the questionnaire it is possible to evaluate the level of knowledge of big data problems. The previous research on the understanding of big data made by the authors was used, where the contents of press articles were analyzed (Wieczorkowski and Polak 2014). On the basis of the problems of big data highlighted in press, particularly related to the social aspect of big data, the list of survey questions was formulated.

The students were chosen as the group of respondents. The students, as young people, are usually open for using new technologies as an obvious part of their lives. As indicated by the above-mentioned studies (Sheehan 2002), such an age group is more diverse in opinions on privacy and moreover students as currently being educated on the higher level should be characterized by sufficient criticism of the described issues. Certainly, this is not a representative sample for the society as a whole. However this choice of respondents allows for posing detailed questions, thanks to their sufficient understanding of the problem.

Thus since the year 2014 the authors conduct a periodical survey on understanding the term of big data and perception of threat coming from using big data methods. The study group consisted of the students from Warsaw School of Economics, which is the university of economic, business and administration profile. The respondents, being the students of economics university should sufficiently well understand the possibilities of new technology applications and at the same time treat them as typical individual or business users.

The survey was conducted during the classes not related to big data, and the questionnaires are in traditional paper form, so the nearly 100% of the issued questionnaires were filled in. The survey is anonymous, but the direct contact with the interviewer improved data reliability. One part of the questionnaire is related to knowledge and understanding of big data term, the second one – perception of threat to privacy. In this paper the authors focus on the second part, referring only to the main conclusions of the first part. The 256 students underwent the survey in consecutive three semesters.

The first part of the survey contained 20 closed questions on the acquaintance with the term big data and pointing its features and related items listed. Because the term is emerging and has multifaceted nature, the authors advocate that it is not possible to unambiguously evaluate the level of knowledge of the definitions of big data. Part of the questions is related first of all to the students' opinions and cannot be used to evaluate their knowledge. The other part refers to the commonly accepted issues and descriptions of big data. The latter questions let the authors conclude on the overall level of knowledge of the big data term among students and a knowledge indicator was created. These issues were comprehensively discussed by the

Authors in the context of IT education within a separate article (Pawełoszek and Wieczorkowski 2015).

The 12 next questions were formulated, referring to the phenomena of mass personal data processing, to which the respondents were asked to assign their subjective perception of threat to their privacy using the scale 1..5. The level 1 meant lack of perception of threat to privacy, the level 5 – feeling of serious threat to their privacy. While choosing questions, the Authors had in mind various threats resulting from mass data processing, in particular related to recent advances in information technologies. The table 1 presents questions from this part along with the arithmetic mean of answers.

Table 1. Questions on perceiving the threat of privacy violence with arithmetic mean of the answers

No.	**Question**	**Arithmetic mean**	**Standard deviations**
1	Using cloud file storage services	**2,7**	1,09
2	Possibility of access by unauthorized persons/robots to private e-mails	**3,6**	1,16
3	Gathering information about users' behavior on the internet (i.e. visiting web pages)	**3,7**	0,98
4	Automatic surveillance of information about activities in social media portals	**3,8**	1,05
5	Gathering data on payments with credit cards	**3,7**	1,15
6	Gathering data on behavior of consumers in loyalty programs	**2,7**	1,08
7	Gathering data on using healthcare services in IT systems	**2,7**	1,19
8	Gathering geolocation data and billings by telecom operators	**3,6**	1,15
9	Gathering data on network usage of devices (i.e. Wi-Fi logs)	**2,9**	1,12
10	Closed-circuit television (CCTV) - video surveillance	**2,6**	1,26
11	Mass photo-taking: aerial, satellite and "street view"	**2,3**	1,21
12	Vehicle identification system i.e. with the aim of charging for the use of roads, detecting traffic offences	**2,7**	1,24

5. INTERPRETATION OF THE SURVEY RESULTS

The average results of answers to the specific questions ranged from 2,3 to 3,8 (with the possible values from 1 to 5). The standard deviations for particular questions ranged from 0,98 (question 3) to 1,26 (question 10). In the light of the above, it is apparent that the respondents answered to the questions in a quite balanced manner, avoiding extreme answers. Generally the feeling of privacy violation was evident because the average evaluation is 3,1.

The strongest perception of threat is represented in case of:

- automatic tracking of activities in social media portals (3.8),
- gathering information on users’ behavior on the internet (3.7),
- gathering information on electronic payments i.e. with credit cards (3.7),
- the possibility of access to private e-mails by unauthorized people/robots (3.6),
- gathering geolocation data and billings by telecom operators (3.6).

These are the threats associated with day-to-day activities, which are inevitable in the modern world.

The slightest perception of threat is related to:

- photo-taking: aerial and satellite imagery, “street view” photos (2.3),
- closed-circuit television (CCTV) - video surveillance (2.6).

These are the activities that the respondents do not have under control.

The quite low perception of threat is related to:

- using file storage services in the cloud (2.7)
- gathering data about consumers’ behavior for the purposes of loyalty programs (2.7).

It can be interpreted as the threats, which could be easily avoided. For example the people taking part in loyalty programs do not treat them as privacy violation, the other people avoid such programs.

One of the questions regarded the sensitive data (health). Interestingly the feeling of threat resulting from gathering data on healthcare is not evaluated highly (2.7). This can be related to legislation, which significantly influences the possibility of processing sensitive data. Moreover the respondents were young people, who usually do not have severe health problems.

The differentiation of the answers is not high for particular questions, but the highest standard deviation can be observed with regard to the situations which are evaluated as the slightest privacy violation: closed-circuit television (1.26), vehicle identification systems (1.24), aerial, satellite and “street view” photo-taking (1.21). The evaluation of such activities is rather ambiguous. On the other hand the lowest differentiation of answers is in the case of one of the activities which causes the largest privacy violation – gathering information on the users’ behavior on the internet, with standard deviation 0.98.

It is worth to notice that the differentiation of evaluations can be seen also over time. Notwithstanding the fact that the period of survey is not long, because it encompasses three consecutive semesters, there was the steady growth of knowledge about big data. Simultaneously there is almost no difference in evaluation of the level of privacy violation. Average evaluations in subsequent semesters are: 3.08, 3.09, 3.09. In case of particular questions any clear trends of change of evaluations cannot be seen.

It seems reasonable to differentiate the answers according to gender, some situations of threat can be perceived differently. There were 144 women and 112 men among the respondents. No significant differentiation can be observed of average total value of feeling threat to privacy according to gender (women 3.09, men 3.07). Interesting differences are in the answers to particular questions. Men clearly feel more significant privacy violation referred to vehicle identification systems (men 2.95, women 2.54) as well as closed-circuit television (men 2.86, women 2.43). These are the same questions for which the bigger standard deviation of answers were observed. The explanation can probably the fact that man are more often caught by the mass surveillance systems in the act of aggressive behavior and road traffic offences. Women perceive higher privacy violation related to data gathering about consumers’ behavior in loyalty programs (women 2.81, man 2.55). It could be result of the fact that women probably pay more attention to loyalty programs, at the same time considering their threats. Women also feel higher privacy violence related to the possibility of accessing private e-mails by unauthorized persons/robots (women 2.75, men 2.36). However it is hard to explain this regularity.

The correlation testing between the indicator of knowledge about big data (obtained from the first part of the questionnaire) and the sense of privacy threat in particular areas may lead to potentially ambiguous results. On one hand the deeper knowledge of big data means greater awareness of potential threats, on the other hand better understanding may raise less concern. The results of correlation tests are ambiguous indeed. Total correlation is not large (for average of all the questions 0.09), for particular questions most often it takes value near to 0 (from -0.03 to +0.09). In case of two questions it exceeds the aforementioned range. These are the questions about possibility of unauthorized e-mail access by persons/ robots (correlation with knowledge is 0.15, one-tailed P value: 0.008) and about automatic tracking the users' activities in social media portals (correlation 0.14, one-tailed P value: 0.012). It can be assumed and confirmed that knowledge about big data methods has very little influence on increasing concerns over unauthorized e-mail access and tracking the internet users' behavior. The justification for this may be the fact that big data methods are based on complex algorithms that are hard to understand for non-IT specialists so many respondents do not have neither the experience nor the ability to assess the risk and threats in this field.

To find out the deeper characteristics of the surveyed group the authors decide to analyze the data using classification trees. Classification trees (or decision trees) are a good choice when the data mining task is classification or prediction of some kind of outcomes. Classification tree labels records and assigns them to discrete classes. Classification trees also provide the measure of confidence that the classification is correct.

The hypothesis of this part of the research was that, there is a difference in perception of threats to privacy regarding knowledge of big data concept. The general characteristics of the surveyed population according to knowledge of big data and their perception of threats in the investigated areas is presented on figure 1.

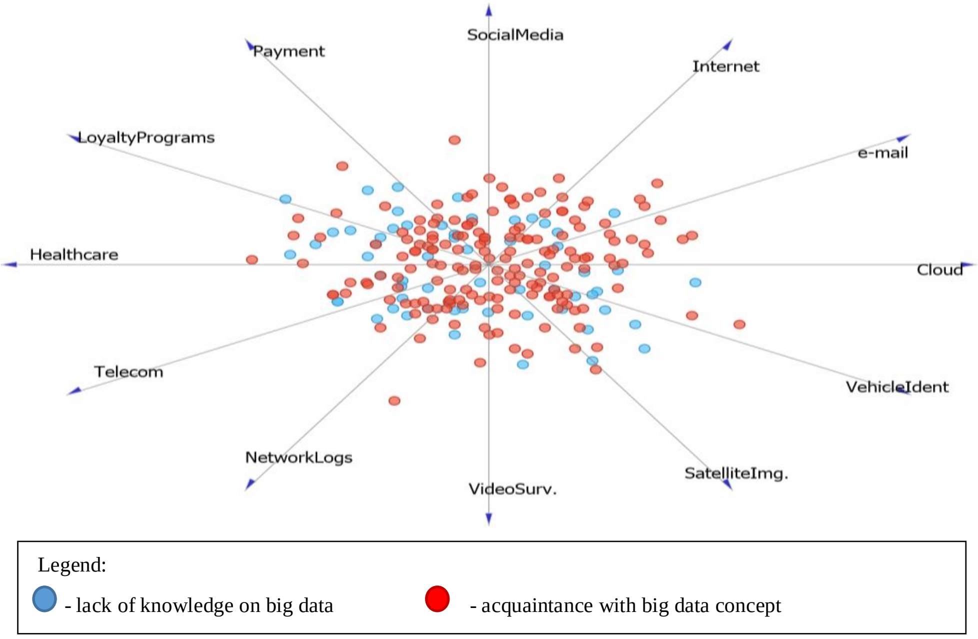

Figure 1. General characteristics of the answers

As it can be seen observations are rather concentrated, however more dispersed are the data characterizing people who have some acquaintance of big data. To describe the population of students and their concerns with regard to privacy the classification tree algorithm were chosen with the following parameters:

- Target Variable – BD – knowledge of big data
- Target class – 1 – acquainted with big data before.

The algorithm was run on Orange 3.3 platform (Demsar et al 2013). Target variable can retrieve data from its target class type. So we choose to describe the group of students that claim some acquaintance with big data. The generated classification tree is presented on figure 2.

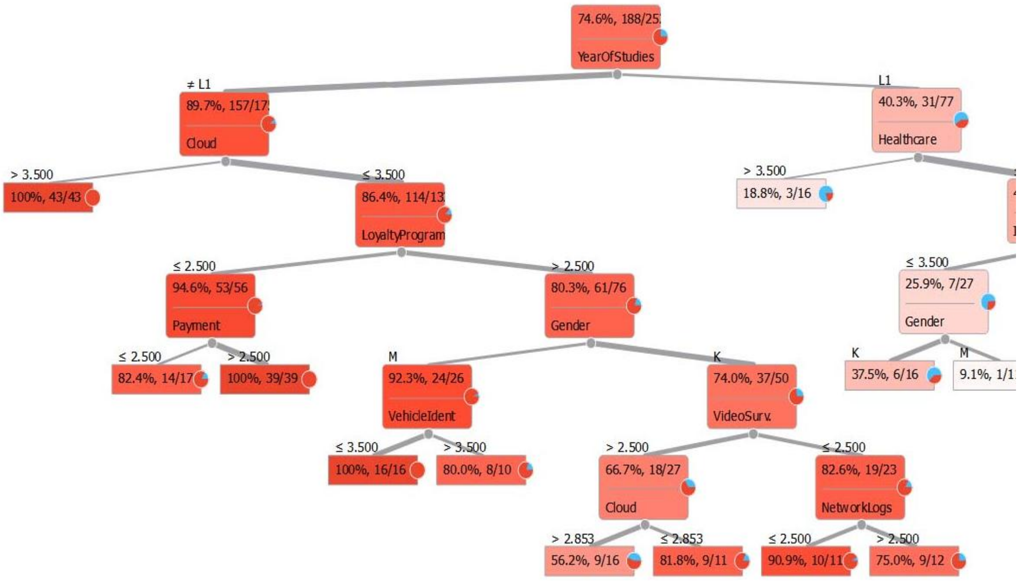

Figure 2. Fragment of classification tree describing the group which has some knowledge of Big Data

As it can be seen most of the students who claim to know big data are in their second or higher year of education - ≠ L1 (157 persons). The important feature is P1 (storing private files on the cloud) – most of the students in this group (114) are not afraid to keep their data on the cloud. The next feature indicated by classification tree algorithm is P6 (gathering data in loyalty programs). Among the persons who fear to share their data in loyalty programs there are 24 men (M) and 37 women (K), however, as the classification tree shows, the level of confidence is higher in case of men. Among the persons who are not afraid to share their data in loyalty programs (53) the majority (39) are also afraid of gathering data on payment cards. The right side of the classification tree is statistically less important as the confidence levels are lower.

The algorithm of classification tree was run one more time with changed parameters to describe the group of students who had not known the big data concept before. The selected parameters of classification tree were:

- Target Variable – BD – knowledge of big data
- Target class – 0 – lack of knowledge

Here the results were statistically less important as the percentage values of confidence show on the classification tree diagram (figure 3). Most of the people in this group (46) was in their first year of study. The most important features were P7 (gathering data on using healthcare services in IT systems) and P3 (gathering information about users' behavior on the internet). Most of the students in this group (33) are not afraid of collecting their medical data by healthcare institutions. Most of those persons are also not concerned with gathering the data about their behavior on the Internet (20). In this group there is equal proportion of men and woman (10).

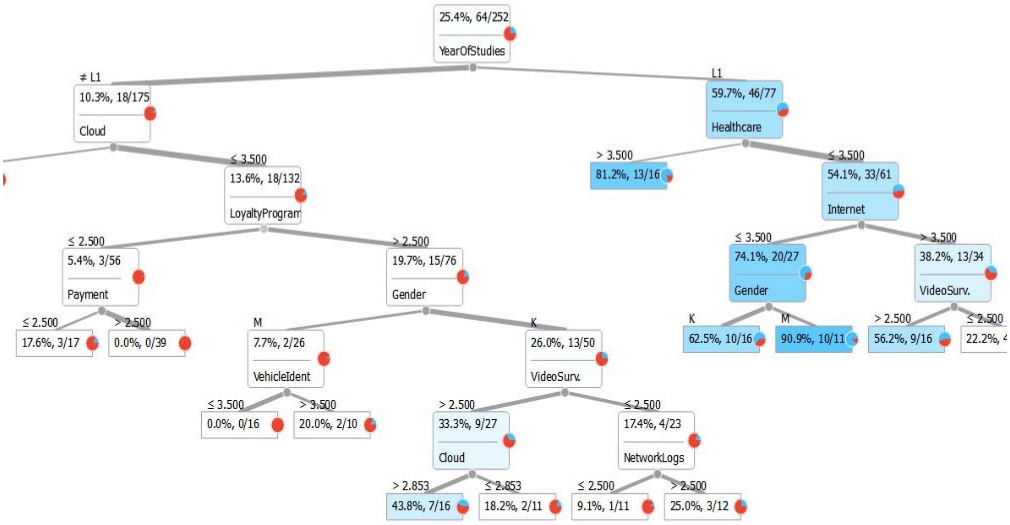

Figure 3. The fragment of classification tree describing group of students who do have not known the big data concept before

6. CONCLUSION

The notion of privacy evolves along with the information technology development, particularly with the concept and practices known as big data. The issues of information privacy are classified into the social aspect of big data (among other aspects: technological and business). It may be assumed that technological developments and business possibilities of big data applications will entail changes in awareness related to privacy of mass processing of personal data as well for commercial purposes (such as personalization of advertisements) as for public management (preventing of frauds and abuses).

The survey conducted by the authors was to identify the level of perception of privacy violation related to different activities associated with big data. The results do not show clear correlation of threats to privacy with the overall level of knowledge on big data phenomenon.

However classification tree analysis identified that different threats are important to respondents regarding the level of their knowledge of big data.

The respondents indicated the activities being to a high degree related with privacy concerns, these were: activities on the internet (i.e. automatic surveillance and gathering data about visited web pages, behavior in social media portals and unauthorized access to private e-mail correspondence), gathering data of cellular telephony (geolocation data and billings) and performing electronic payments (i.e. gathering data about payments with credit cards).

Therefore it is particularly interesting that some of the mentioned activities are totally legal (such as surveillance of the internet users with the aim to personalize advertisement displayed), as well as those which are illegal or officially permitted only in specific cases (for example the activities of authorized forces with the aim to prevent security threats, which encompass the control of private e-mail messages, billings and electronic payments). In turn the lowest perception of privacy violation is connected with activities such as mass photo-taking such as: city surveillance, vehicle identification. Also, interestingly, the low perception of privacy violations is related to processing typical sensitive data – such as using healthcare services. The difference between men and women in total perception of threat related to personal data processing could not be stated although there are clear differences in answers for particular questions. It can be assumed that trust to the legal system and respect for the law influence a person's feelings concerning privacy violation or the lack of it. Therefore the authors advocate that it is very important that the legal acts on the protection of personal data would keep up with the pace of technological development. The law should be on one hand sufficiently detailed, but on the other hand so general and timeless that the technology development would not cause permanent legal gaps. Simultaneously it is very important the law would be actually respected and the society should not be surprised at the news about completely illegal actions of, for example, special forces.

The described research should be continued with the purpose to monitor changes in subjective evaluation of threats to privacy being related to technology advances. The authors are planning to conduct more detailed research considering the aims for which the respondents are able to agree to share their personal data. The question still stays open and the further research can be conducted regarding the differences between regions, culture and legal system of the country influencing the perception of threats to privacy.

REFERENCES

Boyd D., Crawford K., 2012. Critical questions for big data in Information. *Communication & Society*, Volume 15, Issue 5, pp. 662-679.

Craig T., Ludloff M.E., 2011. *Privacy and Big Data*, O'Reilly Media, Sebastopol.

Davenport T.H., Dyché J., 2013. *Big Data in Big Companies*. International Institute for Analytics, http://www.sas.com/resources/asset/Big-Data-in-Big-Companies-Executive-Summary.pdf.

Demsar J. et al, 2013. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* 14 (Aug), pp.2349−2353.

Dumbil E., 2012. *What is big data? An introduction to the big data landscape.*, http://radar.oreilly.com/2012/01/what-is-big-data.html.

GUS, 2013. *Systemy informacyjne administracji publicznej – źródła danych dla badań statystyki publicznej*. Główny Urząd Statystyczny, Warszawa.

Haire A.J., Mayer-Schönberger V., 2014. *Big Data - Opportunity or Threat?,* ITU, https://www.itu.int/en/ITU-D/Conferences/GSR/Documents/GSR2014/Discussion%20papers%20and%20presentations%20-%20GSR14/Session3_GSR14-DiscussionPaper-BigData.pdf

Kamakshi P., *Survey on Big Data and Related Privacy Issues*. IJRET: International Journal of Research in Engineering and Technology Volume: 03 Issue: 12 | Dec-2014, pp.68-70.

Kołodziejczyk Ł., 2014. *Prywatność w Internecie: postawy i zachowania dotyczące ujawnianiadanych prywatnych w mediach społecznych*, Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

Laney D., 2001. *Application delivery strategies*, META Group, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Marwick A. E., Diaz D. M., Palfrey J., 2010. Youth, Privacy and Reputation, *Harvard Law School Public Law & Legal Theory Working Paper Series*, Paper No. 10-29.

McKinsey Global Institute, 2011. *Big data: The next frontier for innovation, competition, and productivity*. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Microsoft, 2013. http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/04/15/the-big-bang-how-the-big-data-explosion-is-changing-the-world.aspx.

Nissenbaum H., 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Nissenbaum H., 2014. Privacy as Contextual Integrity, *Washington Law Review,* 79, pp. 101-139.

Oracle, 2013. *White Paper—Big Data for the Enterprise*. http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf.

Pawełoszek I, Wieczorkowski J., 2015. Big data as a business opportunity: an Educational Perspective, *Annals of Computer Science and Information Systems*, Volume 5, Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, Polskie Towarzystwo Informatyczne, IEEE Computer Society Press, pp.1563-1568 https://fedcsis.org/proceedings/2015/pliks/365.pdf

Pawełoszek-Korek I., 2008. Zastosowanie eksploracji danych w celu pozyskiwania wiedzy z Internetu, *Fenomen Internetu*, Tom II, Wydawnictwo Hogben, pp. 553-559.

PcMag Encyclopedia, 2016. http://www.pcmag.com/encyclopedia/term/ 62849/big-data.

Polańska K, Wassilew A, 2015. Analizy big data w serwisach społecznościowych, *Nierówności społeczne a wzrost gospodarczy*, 4/2015 cz.2, pp. 117-128.

Rouse M., 2011. *Big data*. http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data.

SAS, 2016. http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Sheehan K. B., 2002. Toward a Typology of Internet Users and Online Privacy Concerns. *The Information Society*, 18, pp. 21-32.

Simon P., 2013. *Too big to ignore – The business case for big data*. John Wiley & Sons, Hoboken NJ.

Smith H. J., 1994, *Managing Privacy: Information Technology and Corporate America*, University of North Carolina Press.

Solove D. J., 2002. Conceptualizing Privacy, *California Law Review,* Volume 90, Issue 4, pp. 1087-1155.

Solove D. J., Schwartz P. M., 2009. *Information Privacy Law*, 3rd Edition, Aspen Publishers.

Surma J., 2013. The Privacy Problem in Big Data Applications: An Empirical Study on Facebook, *ASE/IEEE International Conference on Social Computing*, pp. 955-958.

Szymielewicz K., Szumańska M., 2013. *Dostęp państwa do danych użytkowników usług internetowych, Siedem problemów i kilka hipotez*. Fundacja Panoptykon, http://panoptykon.org/files/ panoptykon_dostep_panstwa_do_danych_internet_16.12.2013.pdf.

Tabakow M., Korczak J., Franczyk B., 2014. Big data – definicje, wyzwania i technologie informatyczne. *Business Informatics*, 1 (31), Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, pp. 138-153.

TechAmerica Foundation, 2012. *Demystifying big data: A practical guide to transforming the business of government*, Washington.

Victor N., Lopez D., Abawajy J.H., *Privacy models for big data: a survey*. International Journal of Big Data Intelligence 3 (1), 2016, pp. 61-75

Warren S. D., Brandeis L. D., 1890. The Right to Privacy, *Harvard Law Review*, Vol. IV, No. 5.

Westin A. F., 1996. Privacy in the Workplace: How Well DoesAmerican Law Reflect American Values?, *Chicago-Kent Law Review*, Volume 72, Issue 1, pp. 271-283.

Wieczorkowski J., Polak P., 2014. Big data: Three-aspect approach. *Online Journal of Applied Knowledge Management*, Volume 2, Issue 2, International Institute for Applied Knowledge Management, pp. 182-196. http://www.iiakm.org/ojakm/articles/2014/volume2_2/OJAKM_Volume2_2pp182-196.pdf