Online Journal of Applied Knowledge Management
A Publication of the International Institute for Applied Knowledge Management

Volume 5, Issue 1, 2017

Big data and privacy: The study of privacy invasion acceptance in the world of big data

Jędrzej Wieczorkowski, Warsaw School of Economics, Poland, jedrzej.wieczorkowski@sgh.waw.pl

Przemysław Polak, Warsaw School of Economics, Poland, ppolak@sgh.waw.pl

Abstract

The phenomenon of big data includes technological (new opportunities), business (application), and social aspect. The social aspect applies to the social consequences of the use of big data methods, in particular, those related to the processing of personal and other private data , as well as the danger of privacy violation. In the context of the big data phenomenon, this study presents the results of a survey on the level of acceptance of privacy violation resulting from mass data processing. The different objectives of processing were taken into account, including general, social and commercial. This study helps to draw conclusions concerning commercial and non-commercial use of private data, as well as the legal regulations on personal data processing.

Keywords: Big data, privacy, invasion of privacy, personal data, private data.

Introduction

One of the main directions of information technology (IT) development concerns new solutions designed for processing mass data. It is related to the relatively new concept of 'big data'. It is connected to general phenomena associated with the processing capabilities of large data volumes, the applications of those capabilities, as well as consequences resulting from them.

It is important to properly identify and understand new technological possibilities and their impact on applications in various areas: private, commercial, scientific, government authorities, and public organizations. The applications, in particular those concerning processing of personally identifiable information (PII) and other private data, generate important social consequences. PII is data that can be used on its own or in connection with other information to identify a single person. Private data is data not made available to the general public. An important factor is the evaluation of the possible violation of personal privacy. The practice of using big data solutions, and its formalization in the form of the legal system should, on the one hand, enable the use of new technologies, but, on the other hand, should prevent any form of abuse.

That raises the question about the understanding of privacy in the modern world and the indication of a balance between the efficient functioning of the state or business and the preservation of the rights of individuals. In order to develop that state of equilibrium, it is necessary to understand the needs of the people in relation to their privacy. Those needs are very subjective and change over time along with the availability of new technologies.

This article attempts to answer the question of understanding the privacy needs, relating to the acceptance of and risk associated with privacy violations, by people aware of modern technology capabilities. The research method was based on a survey and a literature review on privacy issues and big data methods. The statistical analysis of the data from the survey was performed using the analysis of arithmetic means, standard deviations, and correlations. The results between subgroups isolated by gender and year of study were also compared.

The concept of privacy is introduced using a traditional approach with an emphasis on the impact of modern technology. The problem of privacy assessment is presented in this study in the context of big data methods. This effect is characterized in relation to the IT evolution. This study presents a view of big data by forming a three-aspect model that includes technological, business, and social aspect. The latter includes the issue of privacy in case of the wide use of such data processing methods.

Theoretical Background

The Concept of Privacy in the Context of Modern Technology

The concept of privacy in comparison to other human rights is relatively young. Privacy is closely linked to the level and style of life, as well as to technological progress. The approach to privacy significantly changed over time and that change was particularly noticeable in the last few years, especially among people with broad access to information and communication technologies (ICT). The concept of 'right to be let alone' developed by Warren and Brandeis (1890) can be considered as the first fairly mature theory of privacy. A timeless element of that theory is the separation of two areas of life: public and private. It was recognized that it is unacceptable to disclose private life, unless it is in the public interest.

Literature distinguishes between the concepts of privacy and the right to privacy. The first determines what privacy is and how it should be assessed. The second concept defines the degree to which privacy should be protected (Solove & Schwartz, 2009). It is important, when analyzing privacy, not to limit oneself solely to a legal approach, because the law can be flawed or can lose touch with reality.

On the one hand, general timeless privacy considerations are still valid. On the other hand, the potential of ICT and, the possibility of automatic mass-processing of private data using the big data method give rise to a completely new way of perceiving privacy. Characteristic for today's possibilities is datafication, which is understood as the passive collecting of diverse data (in particular personal & other private) when processing other data without knowing the purpose of their later use (Mayer-Schonberger & Cukier, 2013). State bodies can collect mass data about citizens not only for current monitoring, but to use those data in the future whenever needed, and sometimes when it will be technically feasible. Data that do not meet the definition of personal data because it does not identify an individual can become personal data when used again due to analyzing and linking together large data sets. So, the methods used for data anonymisation may become ineffective. Consequently, when analyzing the perception of privacy, consideration should now be given to the technical possibilities of mass processing of private and personal data. A factor that may particularly affect the evolution of the current approach to privacy is the use of big data methods.

Currently, it is difficult to find an unambiguous definition of privacy due to its multi-faceted character: e.g. legal, philosophical, sociological, and technological. In the context of ICT, privacy can be understood as control over the flow of private information, in particular on the extent to which the information is made available to others (Westin, 1967). Access control to such information should be considered always in a specific social context - what information, to whom, when and in what situation can be transmitted (Nissenbaum, 2004).

The subject has become a popular topic in publications. According to Kamakshi (2014) the threat to privacy is often perceived by IT users and other persons whose data is processed in information systems. The studies of attitudes towards privacy were undertaken when the Internet and social media were considerably less popular. For example, the Westin (1996) distinguished the following attitudes: fundamentalist (the strongest protection at the cost of giving up benefits), careless (weakest protection), pragmatic (determining in each case the benefits & the costs of sharing private information). The studies by Westin (1996) and Sheehan (2002) showed that at least about half of the population, are pragmatists willing to share their data depending on the benefits. The perspective of obtained benefits explains only to a limited extent the acceptance of privacy invasion for the common good - for public purposes.

The relationship to privacy also depends on the ability to make the critical assessment of one's actions on the Internet resulting from soft barriers related to digital inequality in society. These barriers belong to the so-called secondary digital divisions (primary divisions contain hard infrastructure barriers) and they include, among others, competency and psychological barriers, as well as barriers related to knowledge and awareness of threats (Popiołek, 2016).

Previous studies of social media in the context of the use of privacy settings provide ambiguous results. For example, Surma (2013) argued that the problem is well understood by active users. On the other hand, Kołodziejczyk (2014) showed that students often do not know how to use privacy settings on social networking sites. The approach in such studies focused on actions taken by the users, and therefore, combines the problems of privacy threat acceptance and digital competences, so consequently it is difficult to understand the intentions of respondents. In addition, such studies are limited to risks associated with the processing of personal data placed on the Internet. There are also detailed studies of other problems associated with mass processing of private data, for example in existing smart city solutions (Ståhlbröst, Padyab, Sällström, & Hollosi, 2015), as well as there are attempts to build the general model of privacy issues related to the methods of big data (Victor, Lopez, & Abawajy, 2016).

This study investigates the issue of privacy in terms of an acceptance level for the mass processing of private data for specific (business & public) purposes without limitation to a particular data source. The focus is on the institutional context of privacy in which institutions, including public administration and commercial companies, are potential recipients of private information. Whereas the social context involves individual recipients of information including family and friends (Kołodziejczyk, 2014).

An Outline of Big Data and the Social Aspect

The term 'big data' is not well defined, although such attempts are undertaken. Particularly disputable is forming the term by using the word 'big', which has very general and relative

meaning. The term 'big', in relation to the size of the processed collections of data, has significantly changed its meaning over time, and probably it will change as well in the future.

Various studies on the amount of data collected, even before the analysis of the problem of the size of files that are stored electronically, concerned, for example, the rate of increase in the number of volumes in the libraries (Rider, 1944), an increase in the number of publications and scientific studies (Price, 1975), an increase in the amount of information provided by mass media. The problem was studied in correlation with the size of population and with the number of births. They stressed the fast data growth rate, usually characterized as an exponential increase.

With the development of information technology and information society, the amount of digital data is rapidly increasing. This is due to collecting various data about private individuals, including recording user activity on the Internet, registering services in mobile devices (including location), widespread surveillance in public places, etc. The continuous technological progress, on the one hand, should facilitate searching for required data and organizing information resources, but on the other hand, significantly increases the data volume used in data processing operations. Information overload is a serious challenge for a variety of information systems. In that context, in 1997, the term big data was used for the first time (Cox & Ellsworth, 1997). Due to rapid growth, data volume slowly began to exceed the potential capabilities of storage systems. It was necessary to develop computer database systems able to accommodate available data resources without necessity to dispose of any part of the collected information. Further studies (Bryson, 1999) stressed the importance of analyzes that take place in real time and the ability to extract from data only information, trends and relationships that are useful for the purpose of a particular research.

The basis for the contemporary understanding of the term 'big data' is a 3V model. In the original model derived from the META Group report (Laney, 2001) on the impact of e-commerce, globalization and other trends on the development of information technology, three features were distinguished: volume, velocity and variety. They became the foundation of the big data concept. The volume indicates the large quantity of data being processed, the velocity points to quick changes of data being processed, and the variety represents the different types of processed data. In subsequent years, various authors tried to select other characteristics of big data which could extend the V model: large value of data, veracity of data, data visualization, and visibility of the most relevant data. The growth of the amount of processed data is an evolutionary trend – the result of technological progress and increasing sources of available data. Whereas, the most important features are the big data opportunity of real-time or near real-time processing, and the capability of processing poorly structured data.

Some attempts to define the term consider big data as: the data that exceed the processing capacity of conventional database systems, when the data are too big, move too fast, or do not fit the structures of database architectures (Dumbil, 2012), or data sets whose size exceeds the capacity of conventional database tools for gathering, storing, managing and analyzing data (McKinsey Global Institute, 2011). Big data term is referred to describe the voluminous amount of unstructured and semi-structured data. This study perceives the problem of big data in a broad manner to include three aspects: technological, business and social (Wieczorkowski & Polak, 2014).

The technological aspect includes not only IT capabilities that are widely used in the processing of mass data (High-Performance Computing, in-memory processing, in-database processing, MapReduce paradigm, NoSQL database, etc.), but also possibilities arising from data analysis methods (statistical methods, artificial intelligence, machine learning, data mining, etc.).

Technology development and, consequently, a significant reduction in the cost of processing mass data enabled new applications using such data in business, as well as in the state governance. That aspect of big data was defined as the business aspect. Particularly important is the secondary use of data. The data, irrespective of its initial use, can be reused usually for testing a variety of correlation between them. Often for this purpose, the data sets, which originally were created for completely different purposes, are merged.

A typical application of big data processing is the creation of a personalized advertising message. For this purpose, user activity on the web is monitored and analyzed. Then, the ads watched by them may be based on data profiled in real time. The basic business model of social networking services relies on providing a useful platform in exchange for access to personalized information streams co-written and shared by the community (Polańska & Wassilew, 2015). Personal location data, for example from cellular telecommunication companies, have tremendous business value.

Other examples of business areas for big data applications may include, among others: finance, energy, health service, car fleet management. All the above mentioned applications involve the processing of data related directly to an individual person. Recording and processing financial transactions made with payment cards, e.g. for the needs of the real-time detection of potential abuses, are associated with the processing of information about shopping and location of individuals. The control of a power grid and power consumption in real time in order to manage the entire network results in the processing of data on the behavior of individuals (turning on energy consuming devices). The computer analysis of detailed medical data involves the processing of sensitive data concerning the health of individual patients. The analysis of traffic with automatic recording of the vehicle location, for example in fleet management, is also related to the processing of data about the behavior and location of individuals.

Apart from business applications, personal details are collected and processed by various public institutions to ensure the security of the state. They derive data openly available (Closed Circuit Television (CCTV), aerial photographs and satellite imagery, public Internet content, etc.) and private data obtained on the basis of special rights of various institutions such as the police and special services (data from the telephony: location and billings; private Internet content: e-mails, various files stored in the cloud, etc.). It is possible that some data are obtained by those institutions illegally.

The aforementioned problems of personal and private data processing were classified to the social aspect. Its most important current issue is to ensure the adequate level of privacy. Personal data have been processed for a long time, but the last few years brought such wide possibilities of big data for handling individual personal data, and not only aggregated data on society. Currently, an access to individual data is possible often in real time or close to it, without a delay required for pre-processing data.

Research Method

In this study, an exploratory research was conducted on the problems of big data in its social aspect since 2014. The dilemma is the choice of a methodological approach quantitative or qualitative. The first approach, usually based on surveys of large samples of respondents, requires to determine whether a phenomenon exists, and to determine the scale of its occurrence. It is possible, for example, to study the correlation between the data, but the determination of causality is usually problematic. Qualitative research, using for example observations, individual in-depth interviews, focus group interviews, etc., are not intended to analyze the scale of the phenomenon but its complexity, causes and consequences.

At the current stage, due to its exploratory character, the survey method was chosen. It requires in advance the indication of various activities and purposes of the private data processing associated with the methods of big data. The set of closed-ended questions, which are often used in quantitative research, was considered the most appropriate because of the breadth and multifaceted notion of big data. The responses adopted the classic 5-point Likert scale. In the process of formulating relevant questions were used the results of previous studies of the authors on common understanding of the concept of big data by analyzing the content of newspaper articles (Wieczorkowski & Polak, 2014). Wieczorkowski and Polak (2014) aimed to identify the common understanding of the term big data and related problems, based on media discourse. Consequently, a set of questions was prepared about the subjective sense of privacy associated with various cases of private data processing using big data methods, as well as a set of questions about the level of acceptance of the privacy violation associated with the various objectives of data processing. This paper contains the analysis of the second set of questions.

The respondents of the survey were university students. According to Sheehan (2002), that age group is quite diverse in its opinions on privacy, what enables to observe various dependencies. Students should understand the possible applications of modern technology and should have sufficient criticism of the issue in question. However, the homogeneity of the group rises a problem. It is not representative for the general public. But the selection of respondents, due to sufficient understanding of the issues, allows to ask quite detailed questions. Originally the respondents were undergraduate students of the Warsaw School of Economics, a prestigious state university focused on economic, social and management fields. Later, the respondents included students from two other tertiary education institutions from Poland: a general (Jagiellonian University in Kraków) and a technical university (Opole University of Technology). The study involved 432 respondents in five consecutive semesters.

The survey was anonymous using paper questionnaires and was conducted during classes (not associated with the topic of big data), to achieve the high level of external validity. The main part of the survey covered the feelings about privacy and the acceptance of privacy violation resulting from processing mass data. In parallel, during the first three semesters a survey was conducted on the knowledge and understanding of the term big data (N=256) (Pawełoszek & Wieczorkowski, 2015). Combining the results of those surveys allowed to assign respondents to the general understanding levels of the concept, referred to as an indicator of knowledge.

This study is focused on the issue of privacy violation acceptance. Nine questions were formulated concerning processing mass data, in a situation that could result in a subjective

feeling of privacy violation. Respondents indicated, on a scale of 1 to 5, their level of acceptance of privacy violation for different purposes the processing. Level 1 means the lack of agreement on such an invasion of privacy, level 5 is total acceptance.

The questions were formulated by identifying the different sources of data and methods of data processing: CCTV and traffic monitoring, analyses of billing and geolocation data, controlling e-mails and files stored in the cloud, analyses of public content on the web, analyses of transactional data, e.g. shopping, medical data, analysis of the behavior of Internet users. The list of applications includes various purposes of the processing associated with public and traffic safety, state finances, health care and business marketing objectives.

Results

The average values for the results are within the range from 2.2 to 4.0 (the possible answer values are of 1 to 5, N=432). The results indicate the assessments of the acceptance of privacy violations for different purposes differ quite clearly. The standard deviations for each of the questions are in the range from 1.00 to 1.28. In general, the violations of privacy are clearly noticeable, with an average rating is 3.1 (See Table 1).

The highest permission for violation of privacy concerns for ensuring public safety (e.g. monitoring in public places)' – arithmetic has an arithmetic mean value of 4.0. Probably, it is represents the acceptance of the situation existing for several years (a significant part of life for current students), which is now regarded as something almost obvious. In that case, the violations of privacy are completely open, surveillance cameras are visible, moreover, a statement is often displayed that the area is monitored by CCTV.

Other results above average are for the questions related to the general public safety: the detection of crimes (e.g. the analysis of telephone billing & geolocation data) with a result of 3.6, and counter-terrorism (e.g. the partial control of e-mail & files stored in the cloud) with a result of 3.3. Also, a result highly above average with 3.6 is for improving the functioning of health services and the epidemiological risk prevention, while ensuring anonymization of data about the health of patients (e.g. access to treatment history). This refers to the anonymized data, which, when properly rendered anonymous, secure the lack of access to individual data which is likely to be a reason for the high result.

Table 1. The approval for violations of privacy associated with different objectives - arithmetic mean and standard deviation

No.	To accomplish which of the following purposes would you agree to the violation of your privacy on a scale of 1 to 5?	Arithmetic mean	Standard deviation
1	ensuring public safety (e.g. monitoring in public places)	**4.0**	1.00
2	the detection of crimes (e.g. the analysis of telephone billing and geolocation data)	**3.6**	1.05
3	counter-terrorism (e.g. the partial control of e-mail and files stored in the cloud)	**3.3**	1.28
4	detecting tax violations (e.g. the detection of gray market and tracking assets using the public Internet content	**2.8**	1.19
5	improving transport safety (e.g. traffic monitoring, speed cameras, etc.)	**3.2**	1.24
6	improving the functioning of health services and the epidemiological risk prevention, while ensuring anonymisation of data about the health of patients (e.g. access to treatment history)	**3.6**	1.17
7	preparing customized commercial offers (e.g. for the participants of loyalty programs using the analysis of previous purchases)	**2.7**	1.26
8	the individualization of advertising content (e.g. online advertising displayed in connection with the activity of a particular user or his detected location)	**2.2**	1.17
9	personalized health service offers (using data on the health of patients)	**2.6**	1.26
	Total arithmetic mean	**3.1**	**1.18**

By far, the lowest approval of the privacy violation with a result of 2.2 is for individualization of advertising content (e.g. online advertising displayed in connection with the activity of a particular user or his detected location). Also, other questions related to the personalized marketing content provided results were below the average with a result of 2.7 for preparing customized commercial offers (e.g. for the participants of loyalty programs using the analysis of previous purchases), and results of 2.6 for personalized health service offers (using data on the health of patients). Similarly, a low level of acceptance received for detecting tax violations (e.g. the detection of gray market & tracking assets using the public Internet content) with a result of 2.8. A result of 3.2 is close to the average for improving transport safety (e.g. traffic monitoring, speed cameras, etc.).

The results indicate a much higher acceptance for privacy violations associated with various safety and general welfare issues (except for the issue of tax violation) than the personal information used for commercial purposes. Specific attitude to tax issues may be influenced by local conditions. This might be the specifics of Polish political ethics characterized by a quite high acceptance (much higher than in the countries of Western & Northern Europe) of circumventing tax regulations. As a result, even using public data from the Internet to detect tax fraud does not have a high acceptance.

Attention is also drawn by to the results for question 9, as they did not have significantly lower acceptance than the rest of the commercial offer. However, question 9 received the highest value of results within this group of issues. That result may seem surprising, but it must be taken into account that the respondents are young people who usually have limited health problems. For this reason, they can partially ignore the problems of processing sensitive data about health. At

the same time, the acceptance of processing anonymized medical data for the needs of general public (question 6) is significantly higher.

It is worth paying attention to the question of the highest standard deviation of responses. Higher variation of responses may prove controversies surrounding an issue. Question 9 has high standard deviation of 1.26 regarding the use of personal sensitive data on health. Therefore, that problem is perceived very ambiguously. Also, a high standard deviation of 1.28 for question 3 characterizes the issue of privacy invasion in connection with a counter-terrorism policy. In addition, question 3 in comparison with the others on the general public good, received very low average results. It is therefore also a quite controversial subject, probably associated with concerns over the abuse of the right to control private electronic content. In assessing that result, it must also be taken into account that in Poland there has been in recent years no serious terrorist threats that have taken place in many Western Europe countries.

When interpreting the results, it is important to take into account the extent to which the processing objectives described in each question can actually affect the lives of respondents. The large majority of respondents are unlikely to be exposed to significant conflicts with the law that would result in the surveillance of a person, for example tracking using monitoring, or mobile system data and e-mail analysis. However, massive, often random, machine-based analyses of such data based on automated algorithms that detect, for example, suspicious e-mails are perceived in different way. In the case of some questions, the respondents can only feel discomfort of being monitored but without the effect on their real life. A completely different situation occurs in the case of actions resulting, for example, in the actual receipt of an individualized commercial offer. A mid-situation takes place in the case of mass data processing for the detection of tax irregularities or road traffic offenses when there is a significant likelihood that such methods will be used with negative consequences for the surveyed persons. In practice, on the one hand, objectives can be divided into related to the common interest and to the private interest. On the other hand, it is possible to isolate objectives that affect the lives of those surveyed on different levels.

The results were also analyzed by gender of the respondents (See Table 2). The arithmetic means and standard deviations of results for each question in both groups were compared and statistical significance of differences in results were verified using t-test and determining p-value (t-test for 2 independent means, two-tailed hypothesis). As a level of statistical significance was assumed p-value > 0.05. The italics are used to indicate p-values that do not corroborate the statistical significance of the comparison of the two groups for a particular question.

Results indicated a slightly higher general level of the acceptance of privacy violation among females (3.19) versus males (3.05). However, there are stronger differences in case of specific questions. Females had a higher acceptance to an invasion of privacy related to ensuring public safety (e.g. monitoring in public places)' with results of 4.2 compared to the males results of 3.8. Female acceptance is a also higher for improving transport safety (e.g. traffic monitoring, speed cameras, etc.)' with results of 3.4 compared to males results of 3.0. It is possible that those results are related to some aggressive or dangerous behaviors that are more likely to be performed by males. Males are more likely to be "caught" by the city monitoring for aggressive behavior, as well as by control devices on the road for traffic offenses. In general, females have a

higher level of acceptance of privacy violation in almost all general social issues. The medical data are the only exception, but in case of question 6 this difference is not statistically significant.

Table 2. The results of the survey on the acceptance of privacy violation associated with the gender of respondents

	arithmetic mean		standard deviation			
Question	**Females**	**Males**	**Females**	**Males**	**T-Test Value**	**P-Value**
1	4.2	3.8	0,86	1,11	4.09	0.00005
2	3.7	3.4	0,98	1,10	3.38	0.00079
3	3.4	3.2	1,21	1,35	1.98	0.04794
4	2.9	2.6	1,08	1,30	2.55	0.01120
5	3.4	3.0	1,13	1,32	3.84	0.00007
6	3.6	3.7	1,22	1,11	-1.33	*0.09288*
7	2.7	2.7	1,19	1,33	0.02	*0.49275*
8	2.2	2.3	1,10	1,24	-0.75	*0.22716*
9	2.5	2.8	1,19	1,31	-1.98	0.02436
Total	**3.19**	**3.05**				

The results indicated that males were more likely to agree to privacy violation related to personalized health service offers (using data on the health of patients) than females. This suggests that females assign greater importance to the sensitive data on health. In the case of other questions related to commercial purposes, the differences between the genders are very small and statistically insignificant.

The study also includes the comparison of results from first-year students and students of the second and the third year (See Table 3). Second and third year students were not distinguished due to the high individualization of education in those years. Also in this aspect, the arithmetic means and standard deviations for the results for each question in both groups were compared and statistical significance of differences in results was verified using t-test and determining p-value (t-test for 2 independent means, two-tailed hypothesis). As the level of statistical significance was assumed p-value > 0.05. The italics are used to indicate p-values that do not corroborate the statistical significance of the comparison of the two groups for a particular question.

The results show the increasing acceptance of privacy violations along with the level of study. The differences, however, are small. The statistically significant differences are related to questions about general social goals.

It can be interpreted, that more mature students may better understand the need for the approval to reduce privacy for general public purposes, in particular to ensure common safety. This is shown by the largest increase of acceptance for improving transport safety (e.g. traffic monitoring, speed cameras, etc.) with results ranging from 2.9 to 3.3. In addition to improving

the functioning of health services and the epidemiological risk prevention, while ensuring anonymization of data about the health of patients' with results ranging from 3.3 to 3.7.

Table 3. The results of the survey on the acceptance of privacy violation associated with the year of study.

	arithmetic mean		standard deviation			
Question	**1st year**	**2nd and 3rd year**	**1st year**	**2nd and 3rd year**	**T-Test Value**	**P-Value**
1	3.8	4.1	1.13	0.96	-2.46	0.00714
2	3.5	3.6	1.05	1.05	-0.91	*0.18289*
3	3.2	3.3	1.31	1.27	-1.16	*0.12272*
4	2.7	2.8	1.14	1.20	-0.60	*0.27572*
5	2.9	3.3	1.28	1.22	-2.64	0.00435
6	3.3	3.7	1.15	1.16	-2.85	0.00233
7	2.5	2.7	1.17	1.27	-1.37	*0.08610*
8	2.0	2.2	1.11	1.17	-1.42	*0.07781*
9	2.5	2.7	1.33	1.24	-0.92	*0.18009*
Total	**2.93**	**3.16**				

So far, the duration of the research is limited to five subsequent semesters. However, the authors compared the results from those semesters. But for most questions, no trend in the level of the privacy violation acceptance was identified.

However, an interesting observation is the gradual decline (mostly in 2016) in the approval of question 3 for counter-terrorism (e.g. the partial control of e-mail and files stored in the cloud). This can be interpreted as a result of widely discussed in Polish media anti-terrorism law changes which significantly increase powers of the special services and the police leading to consequently emerging concerns about the possibility of the abuse of that right.

In practice, it can be assumed that the changes of the acceptance level over time may be affected by short-term phenomena, such as published regularly press articles addressing the privacy issues and threats associated with legal changes or the waves of terrorism. However, the differences in results and the rather short period of study do not allow to reach unequivocal conclusions on a significant trend.

The authors also attempted to investigate the correlation between the results to different questions (on a sample of 432 respondents) confirming the theses by calculating P-Value. It confirms intuitively identified groups of questions for which the level of acceptance of privacy violation was similarly assessed. One group includes questions 1-3 on ensuring public safety. The correlation coefficients for the results to a particular pair of questions in this group are in the range from +0.34 to +0.59 (P-value < 0.001). The second distinct group consists of questions 7-9 concerning the use for sales and marketing purposes. The correlation coefficients for results to specific pairs of questions in the group are in the range from +0.42 to +0.60 (P-value < 0.001).

Between questions from both of those groups, the correlation is very low (from -0.11 to +0.09) and statistically insignificant. It shows that the consent to privacy violation for general public purposes does not relate to such approval for commercial purposes.

It is interesting to examine the correlation between the understanding of the problems of big data and the acceptance of the privacy violation. Such relationship is difficult to predict. On the one hand, greater knowledge about big data means greater awareness of various threats, on the other hand, better understanding of the issues can convert into smaller concerns. In order to examine the relationship, the indicator of knowledge of the big data problems formulated for each respondent based on the additional survey was correlated with results to the questions (on a sample of 256 respondents). In general, the correlation is low and statistically insignificant for all the questions (+0.09, P-value=0.151) and for individual questions values are mostly close to zero (from -0.04 to +0.12). The strongest correlation (+0.12, P-value=0.055) occurs in the case of question 6 for improving the functioning of health services and the epidemiological risk prevention while ensuring anonymization of data about the health of patients (e.g. access to treatment history). In that case, it is probably a question of correct understanding of the issues of anonymization and aggregation. A similar result applies to questions 5 for improving transport safety (e.g. traffic monitoring, speed cameras, etc.). However, in that case, it is difficult to explain the reason for the correlation. At the same time, these correlations are so weak that they cannot undoubtedly verify the thesis of the relationship between the understanding of the problems of big data and the acceptance of privacy violation. Nevertheless, despite the difficulty in demonstrating these relationships, it can be expected that the general understanding of the big data phenomenon, i.e. mass data processing and the ability to use them for a variety of social and commercial purposes, has a significant impact on the perception of privacy.

Conclusions and Further Research Plans

The authors distinguished three main aspects regarding the problem of big data. The social aspect was included in addition to the technological aspect (providing opportunities through appropriate technology & analytical methods) and the business aspect (applications in various areas, including commercial and public),. The social aspect includes the social consequences of the use of big data methods especially in regard to the processing of personal and other private data and the subsequent threats to privacy, as well as the issue of legal regulations related to this.

The concept of privacy evolves along with the development information technology. This phenomenon is evident in the review of literature. The awareness of the threats to privacy resulting from the mass processing of personal data is also changing. The study was designed to assess the level of acceptance of privacy violation resulting primarily from processing personal data. This approach extends the studies cited in the article on user privacy settings on the Internet, particularly on social networking sites. This study also includes the sense of privacy violation and its acceptance resulting from the massive processing of personal data and other private data originating from outside the Internet.

There is a clear difference in the level of acceptance between groups of questions covering various purposes of the processing. Significantly higher acceptance of privacy violation occurs in the case of questions related to the needs of general public safety. Clearly lower level of acceptance concerns the questions about the use of personal data for commercial and advertising

purposes. Although this relationship is not clearly correlated with the understanding of big data methods, the respondents probably understand current threat to public safety, and are aware of the effectiveness of preventing them by using the methods of mass processing of personal data. Although, it should be remembered that the respondents were students, who have probably better understanding of these phenomena than an average person.

At the same time, clearly lower levels of acceptance are noted for the processing purposes which can negatively affect the lives of the respondents. This is the way how receiving personalized advertising messages is often perceived. Similar situation concerns the objectives of detecting (and consequently punishing) fairly widespread breaches of law, such as tax avoidance and traffic offenses. The processing of data, which results only in the consciousness of surveillance and does not affect the lives of the majority of respondents, enjoys higher acceptance.

The survey results show that respondents have a fairly high level of trust in the state institutions (a high level of approval for invasion of privacy for common social purposes). Therefore, it seems important to care about that trust by building confidence in the legal system and its practical observance. At the same time, due to the relatively low level of approval for invasion of privacy for commercial purposes. the law should provide sufficient privacy protection in such cases. It is important that the law on the protection of personal and other private data keeps pace with technological progress. It must be sufficiently detailed to hinder breaking the law. On the other hand, it must be, if possible, general so that the development of technology does not cause formation of continuous gaps. The law must be properly balanced between providing the comfort of privacy and ensuring sufficient freedom for business.

Due to the continuous technological progress and changing approaches to privacy the research on the topic should be continued. Further studies are planned to associate the assessment of the privacy violation acceptance to the subjective assessment of the level of privacy violation. Future studies may also involve an international environment in order to identify regional differences. Other research directions may enable the comparison of results for various social groups, such as students of different types of universities, or persons of different age. In order to better understand the studied phenomenon, including causality, future plans are to carry out qualitative research involving individual interviews and in-depth focused group interviews.

Acknowledgement

The authors express gratitude to Malwina Popiołek from the Jagiellonian University, Magdalena Jurczyk-Bunkowska from the Opole University of Technology and Danuta Polak from the Warsaw School of Economics for their help in carrying out the survey. The authors thank also Malwina Popiołek for valuable comments and suggestions.

References

Bryson, S., Kenwright, D., Cox, M., Ellsworth, D., & Haimes, R. (1999). Visually exploring gigabyte data sets in real time. *Communications of the ACM 42*(8), 82-90.

Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. *Proceedings of the 8th conference on Visualization*, 235-244.

Dopierała, R. (2013). *Prywatność w perspektywie zmiany społecznej*. Kraków: Zakład Wydawniczy Nomos.

Dumbil, E. (2012). *What is big data? An introduction to the big data landscape*. Retrieved February 03, 2017 from http://radar.oreilly.com/2012/01/what-is-big-data.html

Kamakshi, P. (2014). Survey on big data and related privacy issues. *International Journal of Research in Engineering and Technology 3*(12). 68-70.

Kołodziejczyk, Ł. (2014). *Prywatność w Internecie: postawy i zachowania dotyczące ujawniania danych prywatnych w mediach społecznych*. Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

Laney, D. (2001). *Application delivery strategies*. META Group.

McKinsey Global Institute (2011). *Big data: The next frontier for innovation. competition. and productivity*. Retrieved February 03, 2017 from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data – A revolution that will transform how we live, work, and think*. Boston, MA: An Eamon Dolan Book / Houghton Mifflin Harcourt.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review 79*, 101-139.

Pawełoszek, I., & Wieczorkowski, J. (2015). Big data as a business opportunity: An Educational Perspective. *Annals of Computer Science and Information Systems 5*, 1563-1568.

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

Popiołek, M. (2016). Nierówności cyfrowe i podziały cyfrowe drugiego rzędu jako wyzwanie dla gospodarki opartej na wiedzy. *Ekonomiczne problemy usług 122*, 115-123.

Price, D. (1975). *Science since Babylon*. New Haven, CT: Yale University Press.

Rider, F. (1944). *The scholar and the future of the research library*. New York, NY: Hadham Press.

Sheehan, K. B. (2002). Toward a typology of Internet users and online privacy concerns. *The Information Society 18*, 21-32.

Solove, D. J., & Schwartz, P. M. (2009). *Information privacy law*. Aspen Publishers.

Ståhlbröst, A., Padyab, A., Sällström, A., & Hollosi, D. (2015). Design of smart city systems from a privacy perspective. *IADIS International Journal on WWW/Internet 13*(1), 1-16.

Surma, J. (2013). The privacy problem in big data applications: An empirical study on Facebook. *ASE/IEEE International Conference on Social Computing*, 955-958.

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

Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review 4*(5), 193–220.

Westin, A. (1967). *Privacy and freedom*. New York, NY: Athenaeum.

Westin, A. F. (1996). Privacy in the workplace: How well does American law reflect American values? *Chicago-Kent Law Review 72*(1), 271-283.

Wieczorkowski, J., & Polak, P. (2014). Big data: Three-aspect approach. *Online Journal of Applied Knowledge Management 2*(2), 182-196.

Authors' Biographies

Jędrzej Wieczorkowski is an assistant professor in the Institute of Information Systems and Digital Economy at the Warsaw School of Economics. He is also an independent IT project consultant and an expert evaluating such projects. His research interests include big data and business intelligence applications and the consequences of using these methods, in particular behavior of IT users and privacy problem.

Przemysław Polak is a senior lecturer and a director of the Postgraduate Studies in Business Analysis in the Institute of Information Systems and Digital Economy at the Warsaw School of Economics. He is also an independent consultant in the field of information systems. His research interests include the methods of modeling user requirements, business process modeling, and the behavior of IT users and decision-makers.