5-2015

Big Data and Privacy: Control and Awareness Aspects

Lucille Perreault
Carleton University, lucille.perreault@carleton.ca

R42. Big Data and Privacy: Control and Awareness Aspects

Lucille Perreault
Carleton University
lucille.perreault@carleton.ca

Abstract

Big data is increasingly used by organizations to be better able to predict consumer behaviour; therefore, allowing organizations to better forecast customer demand and allocate resources accordingly. As big data use increases, a number of questions about the ethical collection and use of consumer data has arisen. For example, the data obtained for use in big data is not always explicitly provided to organizations from individuals. As a result, ownership of the data is not always clear. Organizations must find a balance in exploiting rich customer data and consumers' privacy via its practices and information systems development.

Further research is required on the social implications of big data and the impact to individual privacy. Previous privacy research has indicated the when organizations encroach on an individual's personal boundaries; there is an impact on the relationship between consumers and the organization. Given this, it is imperative that organizations determine best practices on the use and collection of personal data. This paper explores the gap in research on privacy and big data and proposes a research agenda to determine the degree to which the factors of control and awareness account for information privacy concerns with big data and how organizations can utilize these factors to mitigate information privacy concerns.

Keywords

Big data; privacy; control; awareness

1. Introduction

Significant volumes of data are produced daily via social media, mobile phones, web applications, and sensor-abled devices. As globalization and competition increases, companies seek to obtain a competitive advantage. Increasingly, data analytics is being used to assist organizations to forecast demand, and understand customer preferences. Initially, consumer data was limited to what consumers were willing to provide to companies. However, with the onset of more sophisticated information systems, the ability to collect and analyze data has improved. As time progressed, information systems were developed to make "sense" of unstructured data. Data that was previously difficult to obtain and understand has become accessible to organizations with big data and big data analytics. Organizations now have the ability to predict consumer behaviour, without relying on consumers to explicitly provide their information, by scouring the location data on mobile phones, analyzing their web search habits, or their social media posts. As big data use increases, a number of questions about the ethical collection and use of consumer data has arisen. Organizations must find a balance in exploiting rich customer data and consumers' privacy via its practices and information systems development.

Big data and big data analytics can be defined as "data sets and analytical techniques in applications that are so large and complex that they require advanced and unique data storage, management, analysis, and visualization technologies." (Chen et al., 2012). By this definition, big data differs from traditional business analytics by volume, velocity, and variety (McAfee and Brynjolfsson, 2012). In addition, the data collected is generally unstructured. As a result, significant effort is required to make "sense" of the data. When the data is structured and analysed, the data can provide powerful information to assist organizations to create predictive models of consumer behaviour and segment the population to offer customized solutions. In addition, big data has been purported to improve productivity, allow for experimentation and innovation, and measure customer sentiment (Manyika et al., 2011).

The data used in big data, however, is not always explicitly provided to organizations from individuals. Ownership of the data is not always clear. If individuals communicate in the public sphere via social media or enable the location identifiers on their devices, is the information provided now a public good? Ownership becomes more nebulous if another person uploads information on a specific individual. For example, a friend may upload a picture or send birthday wishes on a social media site. Thus far, big data development has assumed that if the information is publicly available, organizations are free to use it. Do organizations have the right to use the data without consent?

In many cases, consumers provide basic information to subscribe to a service, or transact with an organization. If consumers have provided some information, can organizations use that information to collect more insight on consumers? Can organizations use that information for purposes other than what the consumer intended?

As illustrated above, big data generates a number of ethical questions in regards to an individual's right to privacy, and their expectation of the use of their data. Organizations have significant motivation to collect and use consumer data. Organizations believe that big data can provide a competitive advantage via increased knowledge of their consumer. A recent McKinsey report in 2011, for example, states that retail organizations can improve their operating margin by up to 60 percent by utilizing big data (Manyika et al., 2011). With this type of potential, organizations are willing to push the envelope to obtain data that they deem may be critical to their success.

Some recent media accounts highlight some of the ethical issues resulting from big data collection and use. In 2014, OfficeMax sent a mailer to a father in Chicago where the address field included the recent death of his daughter in a car crash (Pearce, 2014). OfficeMax admitted that they obtained the information from a third party provider. In another notorious example in 2012, using a sophisticated algorithm based on shopping habits, Target sent coupons to a teenager for baby items before she had advised her family that she was pregnant (Duhigg, 2012). Despite the bad publicity, Target continues to provide customized coupons, however, Target now includes non-pregnancy related items that can obscure the fact that they suspect the consumer is pregnant (Duhigg, 2012).

Despite these controversial events, research on big data has primarily focused on the artifact and the commercial implications of big data (Pospiech and Felden, 2012). To date, there has been insufficient research on the social implications of big data In contrast, significant research has been conducted in the privacy domain. Some privacy research has been conducted on internet privacy concerns, and consumers' willingness to provide information to obtain some benefit (Hong and Thong 2013, Hann et al. 2002). However, a gap exists in research of information privacy and the collection of "public" data without the knowledge of the participant. This paper explores the gap in research on privacy and big data, and proposes a research agenda to determine the degree to which the factors of control and awareness account for information privacy concerns with big data and how organizations can mitigate information privacy concerns.

Given big data's potential to improve productivity and profitability, organizations may not be willing to change their data use and collection practices as long as there are no significant ethical breaches and its practices comply with regulations. However, previous privacy research has indicated that when organizations encroach on an individual's personal boundaries, there is an impact on the relationship between consumers and the organization (Culnan and Armstrong, 1999, Phelps et al., 2000). Reputational risk and trust not only have an impact on consumers' willingness to transact with organizations, it may result in stricter legislative control.

The primary aim of this paper is to determine how organizations can find a balance in exploiting consumer data, while limiting information privacy concerns resulting from the collection and use of personal consumer data. As a result, an intended outcome of the research is to determine the factors that can cause information privacy concerns, and how these concerns can be mitigated.

This paper will proceed by discussing big data, reviewing the prior research, and evaluating the theoretical foundations of privacy and big data. Based on the foundations of three theories, two constructs of control and awareness will be established to assist in predicting information privacy concerns, and the impact on organizations. Using this theoretical basis, a framework will be created that can assist in developing big data information systems and organizational practices to balance an organization's need for information with individual privacy.

2. Literature review and theoretical basis of research

Given the importance of big data for business and government, academic research in big data has increased. In "Business Intelligence and Analytics: From Big Data to Big Impact," Chen et al. evaluates the evolution of business analytics and big data. The authors categorize the evolution of big data into three distinct time periods, Business Intelligence and Analysis (BI&A) 1.0, 2.0 and 3.0 (Chen et al., 2012). In its first phase, BI&A 1.0 included data analysis and collection by organizations. The data was structured and stored locally in relational databases. As information systems progressed, BI&A 2.0 captured data from the internet and web 2.0 applications. This data is less structured than BI&A 1.0, and requires more data mining algorithm applications. BI&A 3.0 is considered emergent, and focuses on mobile applications and other sensor-abled devices. For the purposes of the analysis of big data and its impact, this study will focus on BI&A 2.0 and 3.0.

In a literature review of big data from 2000 to 2011, Chen et al. found that most publications that included business intelligence and analytics were highly technical (Chen et al., 2012). These articles primarily related to text, data, and web analytics (Chen et al., 2012). Given the complexity of "making sense" of unstructured data, and the speed of analysis required to make real-time decisions, it is understandable that the focus of research has been on the artifact and its potential.

In contrast to big data, significant research has been done in the domain of privacy. In the information systems field, the concept of information privacy is most commonly used when discussing privacy (Belanger and Crossler, 2011). Information privacy is a construct that combines two privacy dimensions of personal communication privacy and data privacy (Clarke, 1999). Personal communication privacy relates to an individual's right to communicate among peers using various media without concern of the communication being collected or monitored (Clarke, 1997). Data privacy represents the assurance that personal data is secure and if that data has been made available to others, the individual can assert some control over the data and its use (Clarke, 1997). Given these two dimensions, it is clear that big data has implications on information privacy. When big data collects social media conversations among peers, it may breach personal communication privacy. Further, if the individual is not aware that their personal information has been collected and has no control of its use, it may break personal data privacy.

Notions of privacy and perceived breaches of privacy differ among individuals. In other words, what would constitute a privacy concern would differ from one person to the next. Researchers have tried to ascertain why these differences exist. Multi-dimensional developmental theory (MDT) asserts that these differences are due to self-ego, environmental, and interpersonal interaction dimensions (Laufer and Wolfe, 1977). Self-ego, or self-development, refers to the development of autonomy and personal dignity within an individual (Laufer and Wolfe, 1977). There are environmental influences, as well, on an individual's view of privacy based on their cultural, social, and physical contexts. The theory asserts that an individual's concept of privacy, based on their environment and self-development, is therefore demonstrated by their interpersonal interactions in a given context (Laufer and Wolfe, 1977). Interpersonal interactions and privacy are characterized by the individual's ability to manage both the interaction and the amount of information provided. As a result, control and choice factor heavily in MDT. An individual can interact with an organization via a website (choice) but may decide not to provide personal information for fear of a breach of privacy (control). MDT has more recently been used in the analysis of the association of interaction and information management with internet privacy concerns (Hong and Thong, 2013).

Once an individual provides their data, they have an expectation that it will be used for its intended purposes. An individual's expectation of the use of their data by organizations can be seen as an implied social contract. Social contract theory has been used as a basis for business ethics (Donaldson and Dunfee, 1994). In general, three elements are considered to be integral to social contract theory: consent of the individual, agreement among moral agents, and a device or method by which an agreement is obtained (Dunfee et al., 1999). Building on the tenets of social contract theory, Donaldson and Dunfee developed the integrative social contract theory (ICST) that outlines that shared norms in an industry act as a foundation for organizations to behave ethically (Donaldson and Dunfee, 1994). In some jurisdictions, by legislation, organizations may be required to have privacy policies in place. An organization's privacy disclosure policy may reduce information privacy concerns and increase the consumer's trust in an organization (Culnan and

Armstrong, 1999). However, in order for policies to be effective, individuals must consent (control) that the data will be collected and understand how their data will be used (awareness). Individuals may initially consent to the collection of their data; however, may not be aware or have any control of their data for “re-use,” resulting in a perceived breach of privacy (Culnan, 1995).

Consumers may willingly provide personal information if they perceive there is some benefit. In information boundary theory, Stanton asserts that an individual’s perception of the relationship with the organization that collects the data, the expected use of the data, and any expected benefit of sharing the data will determine whether an individual is willing to share the data (Stanton, 2003). Individuals control the outflow of personal information through boundary “opening” and “closing” behaviours (Stanton and Stam, 2003). An individual may share information (boundary opening) if there is some benefit, but will withhold information (boundary closing) to mitigate risk.

Information privacy concerns can shape an individual’s trust in an organization (Malhotra et al., 2004). If a consumer’s information privacy concerns are outside their tolerance levels, they may discontinue all future transactions with an organization due to a lack of trust. Consistent with informational boundary theory, consumers may engage in boundary closing behaviours if the risk is deemed too great. Research conducted on trust and e-commerce has found that trust is a mediating factor between information privacy concerns and the willingness to transact online (Van Slyke et al., 2006). In effect, trust is an important component of the decision to transact with an organization; and organizations that lose trust may see a decline in consumer transactions. Due to the potential negative impact, it is important that organizations minimize information privacy concerns when using big data.

Based on the review of the literature, the three theories of multi-dimensional developmental theory, social contract theory, and information boundary theory have common elements of control and awareness in respect to an individual’s concept of information privacy needs.

The degree of awareness and control can determine information privacy concerns; however, the degree may depend on personal privacy risk tolerance. In order to be perceived as being ethical, an organization must ensure that individuals are aware that their data is being collected, and they have control of how their data is used. Building on the three theories, the paper will develop a framework of how organizations should collect and use big data to their benefit, while balancing privacy concerns.

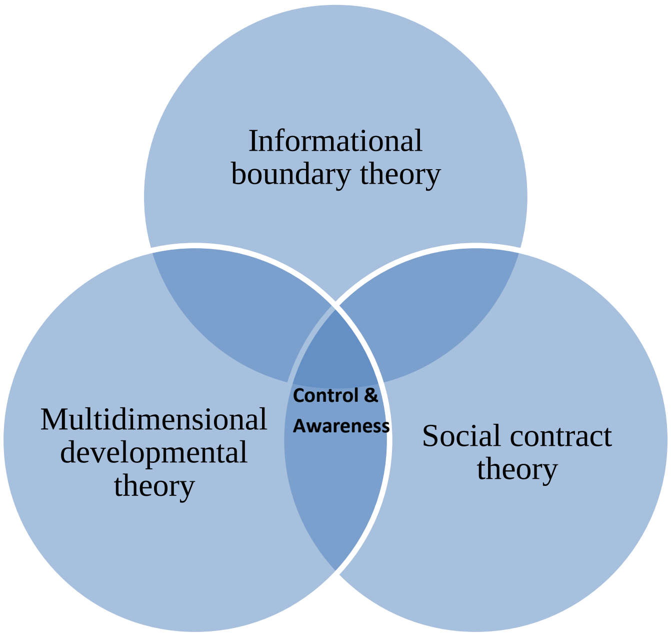
Informational boundary theory
Multidimensional developmental theory
Control & Awareness
Social contract theory

Figure 1: Privacy, Control, and Awareness

3. Research model and hypotheses

Consumer data is generated by social media, transactions in store and online, web browsing, and by sensor-enabled devices. In some cases, consumers agree to provide their data and consent through online terms and conditions. However, these consumers may not be explicitly aware that the data is being collected, nor its intended use. In other cases, the consumer has not consented to data collection, and is not aware that data is being collected. In both these examples, the consumer may feel their privacy has been breached depending on their privacy boundaries.

In the model of the current state in figure 2, consumer data is collected by an organization, and is stored and analysed. Building on privacy research, the awareness, or lack thereof, of personal data collection, and how it will be used, can determine information privacy concerns (Culnan, 1995, Clarke, 1997). Information privacy concerns are also influenced by consumers' ability to control their information (Clarke, 1997). In big data, as consumer data is collected, the two constructs of awareness and control will influence a consumer's information privacy concerns. Similarly, the use of the data can generate information privacy concerns depending on the level of awareness and control a consumer has of their data. In effect, awareness is twofold; awareness can indicate whether a consumer is aware that there information is being collected, and if they are aware on how the data will be used. Control is multifaceted as well. Control can include consent to provide the data, and consent that the data will be used for its intended purpose. Control can also encompass the ability to rescind consent and/or have the data returned or purged. In addition, control could include restriction of re-use or secondary use of personal data. As a result, in the big data domain, consumer awareness and control of the collection of data and its use can determine an individual's information privacy concerns.

H1a: Awareness can influence information privacy concerns.
H1b: Control can influence information privacy concerns.

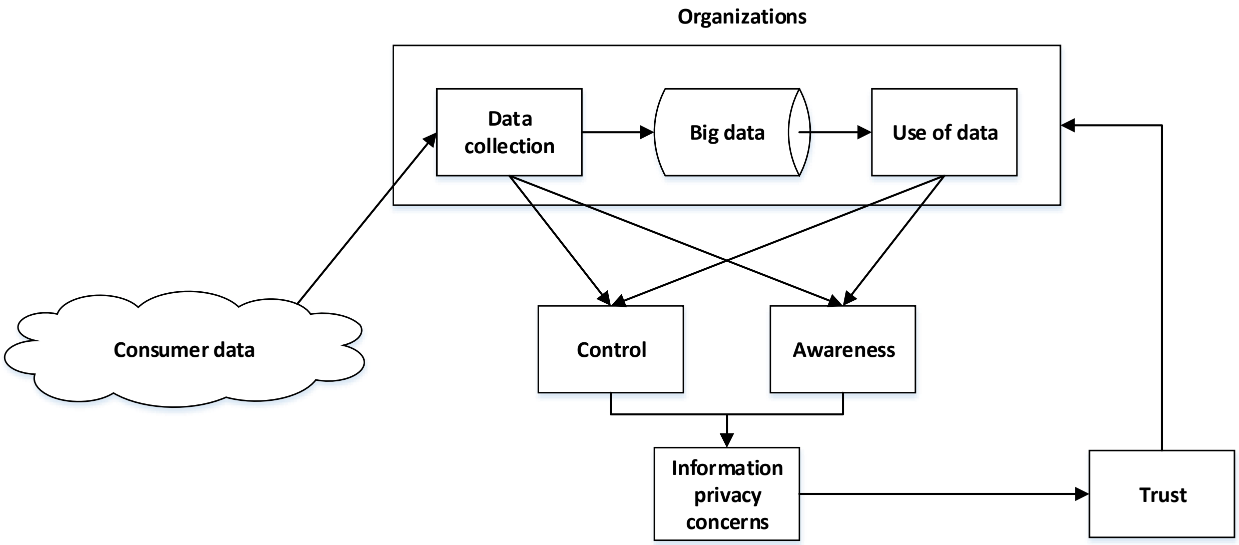

Figure 2: Current state

Consistent with social contract theory, consumers may consider that collection and use of their personal data is only fair when they are advised of the intended use (awareness), and are granted control over the data. Organizations that breach this implied social contract of fairness may no longer trust the organization with its personal data.

H2: Information privacy concerns can influence the feelings of trust for an organization.

Integrative social contract theory (ICST) developed by Donaldson and Dunfee asserts that shared norms in an industry assist in developing a standard on ethical behaviour in a given domain (Donaldson and Dunfee, 1994). Given the potential for collection and use of data without consumer consent, and the impact on information privacy concerns, organizations that collect data from consumers or potential consumers should establish shared norms of data collection and use. An example of a shared norm within an industry includes PCI (payment card industry) data security standards. The standards are upheld by individual payment card companies such as Visa Inc. or MasterCard; however, the standards are aligned among all five founding payment card members.

Given the potential competitive advantage of collecting and using big data, a collective code of conduct among organizations in industry will be required to ensure fairness. If no shared norms can be established, legislative standards may need to be put into place to ensure compliance. In a voluntary or legislative framework, disclosures of data collection, and its use, will improve awareness and reduce information privacy concerns.

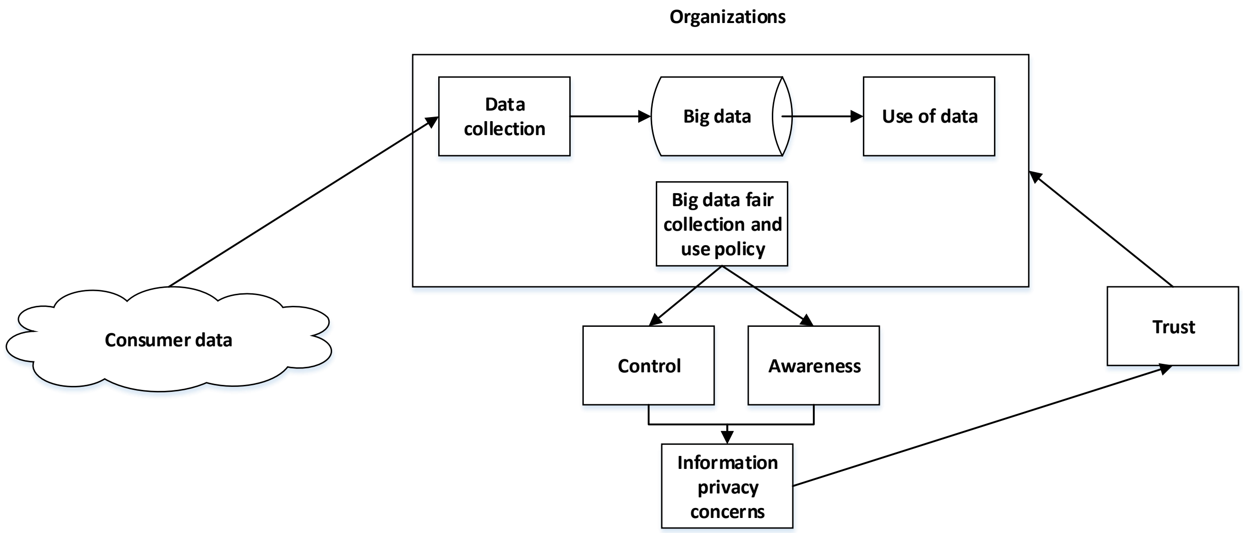

Figure 3: Proposed model

Figure 3 illustrates a proposed model that outlines the impact a big data fair collection and use policy may have on consumers and organizations. A shared norm of a big data disclosure policy of industries that use big data can assist in building awareness of data collection and its use. A fair data collection and use policy should encompass a mechanism to notify consumers of the data collection and its intended use (awareness).

H3a: Transparency of collection and use of data will increase awareness.
H3b: Presence of a big data fair collection and use policy among organizations in an industry will reduce privacy concerns of consumers.

In addition to awareness, consumers require an element of control of the data to reduce information privacy concerns. Consistent with MDT, consumers will have different boundaries of privacy. As a result, different options to control their data should be provide in a big data policy. For example, consumers should have the ability to opt out of data collection. In addition, even if a consumer consents to collection, a consumer should also have the ability to strip any identifiers. Considerable research has been conducted on anonymization, pseudonymization and data masking tools for big data. For example, a field experiment was conducted on smart phones where personal data was not transmitted to central servers, thereby reducing information privacy concerns (Sutanto et al., 2013). Allowing greater control of data collection practices and use of the data will diminish privacy concerns. A big data disclosure policy should outline procedures for consumers to choose to opt out, or choose options to strip any of their identifiers if requested.

H4: A comprehensive big data fair collection and use policy that increases consumer control on data collection and use will reduce information privacy concerns.

Consumers may not necessarily transact with organizations that collect data for analytical purposes; as there may be no expected benefit of an anticipated transaction. However, if awareness and control of the data is within acceptable boundaries for an individual, the information privacy concerns should be diminished. Reduced risk to privacy should increase the perceived trustworthiness of an organization.

H5: Reduced information privacy concerns due to a comprehensive big data collection and use policy will increase trust in organizations that use big data.

4. Research methodology

In order to test the model and the hypotheses, a cross-sectional research design has been selected. Given that perceptions of privacy and privacy breaches vary by individual, a cross sectional research design will be used to measure the constructs among a diverse population. Cross-sectional research has been selected since it less resource intensive. In addition, since big data and privacy is a relatively new area of research, it is important to obtain information in a timely manner. The research design will specifically employ an online survey as the method to measure the constructs of awareness, control, information privacy concerns, and trust. Further studies will be required to confirm the causality of these relationships

Leveraging on research conducted on information privacy concerns, we will define “control” as the degree to which an individual is concerned that they do not have sufficient control of the use or collection of their personal information (Malhotra et al., 2004). “Awareness” can similarly be defined by the degree to which an individual is concerned by the data collection practices and use of personal data by organizations (Malhotra et al., 2004, Culnan, 1995). The measurement of control and awareness will be based on the scale developed by Malhotra et al. as part of e-commerce research and information privacy concerns (Malhotra et al., 2004). Similarly, we will utilize Bhattacherjee’s seven item scale of trust to measure trust in the research model (Bhattacherjee et al., 2002). The seven item scale of trust is based on three dimensions of ability, integrity, and benevolence (Bhattacherjee et al., 2002). Within the ability dimension, expertise and information are two attributes required for an individual to trust an organization. The organization must have the ability (expertise) and the knowledge (information) to do what they say they will do. Within the integrity dimension, the fairness in transaction and service are paramount; whereas, in the benevolence dimension, empathy and resolving concerns are required to build and maintain trust. Further, the scale utilises a measure of overall trust based on the three dimensions used.

The online questionnaire will target users of social media and/or users of location enabled smart phone users. This type of user has been selected for measurement as these users are more likely to have data collected and analysed by big data organizations. Users who do not transact on these platforms may not have the same concerns about data collection. A big data disclosure policy, for example, may not have the same impact if consumers are not “connected.”

The online questionnaire will be sent to a sample population to be able to ascertain their viewpoints of collection of data for use in big data analytics. The questions will be structured on a five point Likert scale ranging from “strongly disagree” to “strongly agree.” The survey will include questions to measure whether control and awareness of big data collection practices influence information privacy concerns. The survey will also measure whether a shared industry standard of big data collection will reduce their information privacy concerns. Subsequently, the relationship between trust and information privacy concerns will be evaluated.

An adequate sampling frame will be required for the analysis. A random sample of users of Linkedin, Twitter, and Facebook will be created. Assuming a five percent response rate, an internet survey will need to be sent to at least 4,000 individuals. Two hundred responses should be sufficient to proceed with a factor analysis of the constructs. Assuming the data will be normally distributed, a factor will be used to determine if there is an association between control and awareness and information privacy concerns. Similarly, the analysis will determine the relationship between information privacy concerns and trust, and the relationship between a big data fair data collection and information privacy concerns.

5. Conclusions and limitations

Some potential limitations of this study should be mentioned. First, the response rate of the online survey may be an issue, and may have an impact on the analysis. A response bias of individuals that are not sensitive to privacy issues may also occur. Respondents, for example, may be more prone to share information than those who do not engage in online surveys.

The measures used in the study were originally devised for trust and privacy concerns in the internet and e-commerce contexts. These measures may not sufficiently robust for measurements of control, awareness, and trust in the BI&A 3.0 context.

However, given the significant ethical issues of big data practices, further research is required to determine if awareness and control influence information privacy concerns and trusting beliefs of organizations that collect and use big data. Further research is also required to evaluate whether big data disclosure policies and tools that allow consumers to be aware of the collection of their data, and also allow control of the use of their data will be sufficient to reduce information privacy concerns. Finding a balance between the potential economic benefit to organizations and consumer privacy concerns is an important area of research.

References

Belanger, F. and Crossler, R. E. (2011). Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems, *MIS Quarterly,* 35(4), pp. 1017-1041.

Bhattacherjee, A. (2002). Individual Trust in Online Firms: Scale Development and Initial Test, *Journal of Management Information Systems*, 19(1), pp. 211-241.

Chen, H., Chiang, R. H. L., and Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact, *MIS Quarterly*, 36(4), pp.1165-1188.

Clarke, R. (1999). Internet Privacy Concerns Confirm the Case for Intervention, *Communications of the ACM*, 42(2), pp. 60-67.

Clarke, R. (1997). Introduction to Dataveillance and Information Privacy and Definitions of Terms, www.rogerclarke.com/DV/Intro.html.

Culnan, M. J. (1995). Consumer Awareness of Name Removal Procedures: Implications for Direct Marketers, *Journal of Direct Marketing*, 9 (2), pp. 10–19.

Culnan, M. J. and Armstrong, P.K. (1999). Information Privacy Concerns, Procedural Fairness, and Impersonal Trust: An Empirical Investigation, *Organization Science*, 10 (1), pp. 104–115.

Donaldson, T. and Dunfee, T. W. (1994).Toward a Unified Conception of Business Ethics: Integrative Social Contracts Theory, *Academy of Management Review*, 19(2), pp. 252-284.

Dunfee, T. W., Smith, N.C., and Ross, W.T. (1999). Social Contracts and Marketing Ethics, *Journal of Marketing*, 63 (3), 14–32.

Duhigg, C. (2012) How Companies Learn your Secrets, *NY Times*, 19 February, p. MM30.

Hann, I. H., Hui, K.L., Lee, T.S.U., and Png, I.P.L. (2002). Online Information Privacy: Measuring the Cost-Benefit Tradeoff, *Proceedings of the 23rd International Conference on Information Systems*, Barcelona, Spain. December 15-18, pp. 1-10.

Hong, W. and Thong, J. (2013). Internet Privacy Concerns: An Integrated Conceptualization and Four Empirical Studies, *MIS Quarterly*, 37(1), pp. 275-298.

Laufer, R. S. and Wolfe, M. (1977). Privacy as a Concept and a Social Issue: A Multidimensional Development Theory, *Journal of Social Issues,* 33(3), pp. 22-42.

Malhotra, N.K., Kim, S. S., and Agarwal, J. (2004). Internet Users' Information Privacy Concerns (IUIPC), Information Systems Research, 15(4), pp. 336–355.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

McAfee, A. and Brynjolfsson, E. (2012). Big Data: The Management Revolution, *Harvard Business Review*, October 2012, pp. 61-68.

Pearce, M. (2014) "OfficeMax Executive Apologizes over 'Daughter Killed' Mailer", *LA Times*, January 20.

Phelps, J., Nowak, G., and Ferrel, E. (2000). Privacy Concerns and Consumer Willingness to Provide Personal Information, *Journal of Public Policy & Marketing*, 19(1), pp. 27-41.

Pospiech, M. and Felden, C. (2012). Big Data - A State-of-the-Art, *AMCIS 2012 Proceedings*, Paper 22, http://aisel.aisnet.org/amcis2012/proceedings/DecisionSupport/22.

Stanton, J.M. (2003). Information Technology and Privacy: A Boundary Management Perspective, *Socio-Technical and Human Cognition Elements of Information Systems, First Edition*, Clarke, S. E. Coakes, G. Hunter, & A. Wenn, London: Idea Group, pp. 79-103.

Stanton, J. M., Stam, K. R. (2003). Information Technology, Privacy, and Power within Organizations: a view from Boundary Theory and Social Exchange perspectives, *Surveillance & Society*, 1(2), pp. 152-190.

Sutanto, J., Palme, E., Tan, C., Phang, C. W. (2013). Addressing the Personalization-Privacy Paradox: An Empirical Assessment from a Field Experiment on Smartphone Users, *MIS Quarterly*, 37(4), pp. 1141-1161.

Van Slyke, C., Shim, J.T., Johnson, R., and Jiang, J. J. (2006). Concern for Information Privacy and Online Consumer Purchasing, *Journal of the Association for Information Systems*, 7(1), pp. 415-442