Journal of Educational Technology Development and Exchange (JETDE)

Volume 8 | Issue 2

Article 5

12-2015

Big Data Ethics in Education: Connecting Practices and Ethical Awareness

Xiaojun Chen

Chen Ying Liu

Recommended Citation

Chen, Xiaojun and Liu, Chen Ying (2015) "Big Data Ethics in Education: Connecting Practices and Ethical Awareness," *Journal of Educational Technology Development and Exchange (JETDE)*: Vol. 8 : Iss. 2 , Article 5.
DOI: 10.18785/jetde.0802.05
Available at: http://aquila.usm.edu/jetde/vol8/iss2/5

Chen, X. & Liu , Y. (2016). Big Data Ethics in Education: Connecting Practices and Ethical Awareness. *Journal of Educational Technology Development and Exchange*, 8(2), 81-98.

Big Data Ethics in Education: Connecting Practices and Ethical Awareness

Xiaojun Chen
St. John's University

Ying Liu
St. John's University

Abstract: *The purpose of this paper is to discuss big data ethics in education. To achieve this goal, this paper first discusses big data from its origin, and then discusses big data ethics from its philosophical perspectives of cyberethics and the emphasis on privacy issues in using big data in researching and teaching. Cases, policies, and code of conduct regarding big data and privacy are discussed with ethical considerations from data ownership and privacy, as well as instructor and learner responsibilities perspectives. Key privacy preserving data mining techniques are also discussed, and the authors recommend using a hybrid approach to address privacy concerns in educational big data context. This discussion aims to broaden the discussion on cultivating researchers' ethical awareness in employing and designing future big data research in education, as well as to raising data analyst' ethical awareness in employing big data in an educational open data context. Future studies in exploring empirical practices to cultivate big data ethics are recommended.*

Keywords: big data, ethics, privacy preserving data mining, policy, code of conduct

1. Introduction: Big data and its origin

With the advancement of technologies, large amounts of data are generated and accumulated when we use different devices to make phone calls, use search online, make purchases through e-commerce web sites, make transactions in a supermarket, read data from sensors, use social media to interact with our friends, or use the Geographical Positioning System (GPS) when traveling in our vehicles. The amount of data humans generate in just 24 hours is equal to 70 times the information held in the Library of Congress (Smolan & Erwitt, 2012). Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze (Manyika et al., 2011).

Although the field of big data research

expanded rapidly in the past several years, attracting researchers' attention from different disciplines, big data is not a completely new concept. When super computers were developed in the 1970s, the concept of a "database machine" emerged (Chen, Mao, & Liu, 2014). The term "big data" was used to refer to a large data set that could be generated, collected, and analyzed by super computers and database machines (Cox & Ellsworth, 1997; Wainer, Gruvaeus, & Blair, 1974). With the advancement of information technology, however, the data generated daily via the Internet and other networks are growing exponentially. This allows more opportunities to use the data being gathered through multiple platforms and devices. Thus, researchers, policy makers, industrial leaders, and educators from different disciplines are attracted to the field of big data research order to investigate and understand big data in the context of their discipline and unique perspectives. See for example such research in the fields of science engineering (Wu, Zhu, Wu, & Ding, 2014), computer science (Jagadish et al., 2014), medical informatics (Bourne, 2014, Vaitsis, Nilsson, & Zary, 2014), neuroscience (Sejnowski, Churchland, & Movshon, 2014), healthcare (Murdoch & Detsky, 2013), business (Chen, Chiang, & Storey, 2012), management (Shepherd & Watters, 2013), economics (Einav & Levin, 2014), sociology (Swlwyn, 2015), and education (Papamitsiou & Economides, 2014).

IBM data scientists break big data down into four dimensions: volume, velocity, variety, and veracity (4 V's) (IBM, n.d.). The dimension volume refers to the scale of the data. From the beginning of recorded time until 2003, we created 5 billion gigabytes (exabytes) of data. In 2011, the same amount was created every two days. In 2013, the same amount of data was created every 10 minutes. The dimension velocity refers to the analysis of streaming data. As data are accumulated every second, data quickly become out-of-date, so it is important to use the data as fast as possible (Liu, 2014). The third dimension, variety, shows different types of data we collect, e.g. structured data, unstructured data, text data, numerical data, image data, audio and video data (Liu, 2014). Veracity means the uncertainty of the data. The collected data may contain noise, while it is unknown which data are accurate and which data are noises (Liu, 2014).

These four dimensions are also applied to educational big data (Baker, 2015). When learners interact with a digital device, data about that interaction can be easily captured or "logged" and made available for subsequent analysis (Baker & Inventado, 2014). Researchers have recently used data from tens of thousands of students (the dimension of volume) (Allen & Seamen, 2013; Baker, 2015). Every student entry on a course assessment, discussion board entry, blog entry, or wiki activity could be recorded, generating thousands of transactions per student per course (the dimensions of variety and veracity) (Picciano, 2012). Furthermore, this data would be collected in real or near real time as it is transacted and then analyzed to suggest courses of action (the dimension of velocity) (Picciano, 2012).

Data in current education research tend to be huge. For example, among education data publicized in data.gov, a student score card datasheet file (.CSV) can contain as much as 120 megabytes of data per year. Any analytics to be done on such a dataset would require more computation power and memory than any personal computers can handle, not to mention analysis over several years. Researches with about half a million records can easily take up to a hundred megabytes of data. In the process of research, temporary data generated can take up several gigabytes of memory and hundreds of hours of computation. Education is increasingly occurring online; as a result, digital learning technologies such as games and online learning systems collect vast amounts of data as the students' progress through the game, test, or activity. In a survey of online learning in 2014 (Allen & Seaman, 2015), five million US students were reported as participating in online learning at some level. Mobile, digital, and online technologies are increasingly utilized in many educational contexts.

The exploration of data can give a broader picture of the learning process than traditional measures such as grades and test scores. It can also help educators and researchers gain valuable insight into how to improve and personalize learning for students, which can be used to improve educational effectiveness and support basic research on teaching and learning (Francisco, n.d.). Data analytics is also used in the areas of enrollment management, student progress, and institutional finance and budgeting (Bichsel, 2012). Nevertheless, those incredible benefits could be jeopardized if we do not address the issues of security and privacy (Payton & Claypoole, 2014), and prevent important societal values from becoming subordinate to the new capabilities of big data. Hence, though the promises of big data aroused researchers' and practitioners' interests in integrating cloud-based platforms in education to store data and access data, it is reported that parents, teachers, and school administrators at US K-12 classrooms are still holding back such practices because of the privacy concerns (Castro, 2013). Such conflicts of interest show that it is important to understand 1) the big data ethics in such a transparency age, as well as 2) policies and techniques that would assist educators and data analysts in educational context to work better with big data.

Therefore, the purpose of this paper is to discuss the big data ethics, from its origins from cyberethics and implications on cultivating big data ethics in education. Figure 1 highlights the framework for discussion in this paper. The authors first discuss the big data ethics from philosophical perspectives of cyberethics and the emphasis on privacy issues in using big data in research and teaching. The discussion continues in key cases and policy related to big data usage in various fields. Key privacy preserving data mining techniques are also discussed, and the authors recommend using a hybrid approach to address privacy concerns in an educational data context.

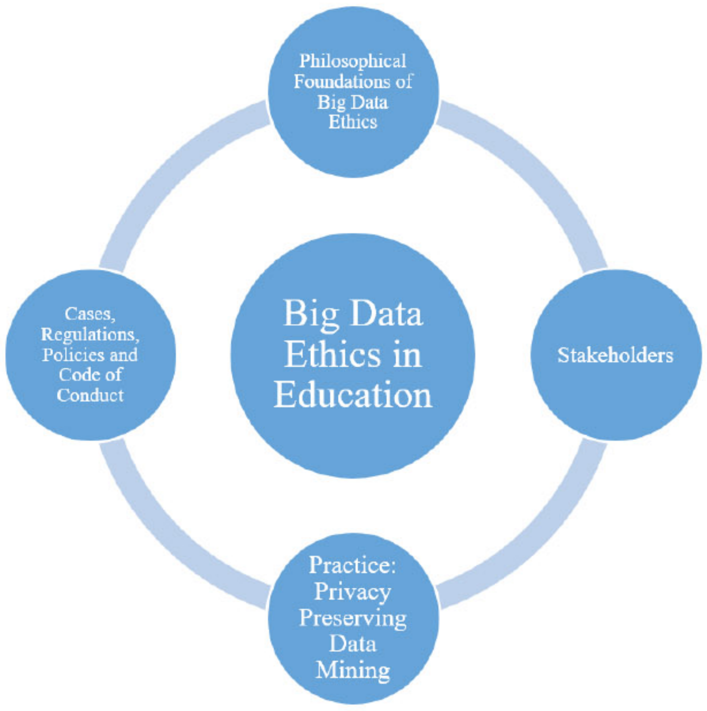

Philosophical Foundations of Big Data Ethics
Big Data Ethics in Education
Cases, Regulations, Policies and Code of Conduct
Stakeholders
Practice: Privacy Preserving Data Mining

Figure 1. Framework for Discussion on Big Data Ethics in Education.

This discussion aims to broaden the discussion on cultivating researchers' ethical awareness in employing and designing future big data research in education, as well as to raising data analyst' ethical awareness in employing big data in educational open data context.

2. Big data ethics: From its origin to current challenge

In this section, the authors discuss the origin of big data ethics in the light of understanding possible ethical dilemmas that would arise in the age of transparency. Though many researchers and practitioners have investigated the initiatives that big data could impact different fields, including education, many scholars and practitioners raised concerns of utilizing big data. In her editorial article, Eynon (2013) called for attention on big data ethics: "Big Data represents a number of ethical considerations, particularly around privacy, informed consent, and protection of harm, and raises wider questions of what kinds of data should be combined and analyzed, and the purposes to which this should be put" (p. 238). Incidents around privacy and publicized data have been discussed among researchers from different countries. Zimmer (2010) examined the research ethics based on privacy issues around social networks, such as Facebook. Yang, Hung, and Lin (2013) critically analyzed a hacker-released database containing a huge amount of website user data from China. Eynon (2013) challenged education researchers to join the debate of what does big data mean for education, specifically with regard to the ethical challenges of privacy and big data analytics.

Big data is often stored and analyzed in cloud-based platforms and systems, thus it is inevitably a part of cyberspace. Therefore, big data ethics are part of the continuing discussion of cyberethics. Norbert Wiener (1948, 1950, & 1964) highlighted in his books several ethical problems related to computers that are still important today, which marked the beginning of the field of cyberethics. A few other ethical terms were not used until other researchers joined the discussion on ethics with varied scopes and focuses, such as computer ethics (Maner, 1980), information ethics (Hauptman, 1988), and Internet ethics and cyberethics (Spinello, 2000). Specifically, the term cyberethics is used when ethical aspects related to computer communicating and networking are being studied (Spinello & Tavani, 2004). Cyberethics examines the moral, legal, and social issues in the use and development of cyber technologies (Tavani, 2010). It studies various ethical issues that arise from the communication of different types of computing devices, ranging from hand-held devices and personal computers to mainframe computers.

It is not clear whether the development of the Internet created ethical issues that were not studied before. Johnson (2001), as oppose to Maner (1996), argued that issues in cyberethics are simply new subclasses of ethical problems that have long existed and no addition to the theory is needed. Johnson (1996) suggested that there are three important special features of the Internet with respect to ethics: the global scope of the Internet, the anonymity of the Internet, and the reproducibility of information. These features may create differences in between online and offline ethical behavior.

As Johnson (2009) has observed, big data requires transparency. In a modern world, transparency enables informed decision. Lastly, big data can compromise identity. Identity is the ability for an individual can define who they are. The development of big data can induce the risk of an individual being identified, categorized, and modulated before the user makes up his/her mind on who he/she is (Davis, 2012; Richard & King, 2014). Major ethical concerns in the field of big data research, as suggested by researchers (Azari, 2003; Schultz, 2006; Quigley, 2007; Baase, 2013; Kizza, 2014), are privacy, anonymity, freedom of speech and censorship, safety, security, intellectual property, digital divide, professional code of ethics, and ethics education. Among those concerns, privacy, security, and professional code of ethics are most frequently studied. In the following section of this paper, the discussion focuses on big data dilemmas, specifically privacy.

3. Cases, policies, and code of conduct- The Big Data Dilemma

In this section, several cases of employing big data are discussed with focus on big data regulating policies. There are successfully implemented projects and initiatives with big data in organizations, as well as setbacks in education and other industries. Discussing these cases brings a clearer picture in understanding the landscape of applying big data analytics in education. The purpose of this discussion is to provide both sides of the story when implementing and using big data in education world. Policies and regulations published around big data ethics are also discussed with consideration to compliance with codes of conduct in this section.

In the book Building a smarter university: Big data, innovation and analytics (Lane, 2014), researchers discussed possible practices of integrating big data in higher education from different perspectives, for example, in college admission practices, recruitment strategies, students successes and course equivalencies, and measuring the internationalization of higher education. Universities and colleges have been using predictive analytics to shape policies and practices to ensure students' success. Purdue University adopted a Course Signals program, which produces red, yellow, and green evaluations of student behaviors in comparison with past behavior of successful students. Course Signals detects early warning signs and provides interventions to students who may not be performing to the best of their abilities before they reach a critical point. Instructors use Course Signals to detect early warning signs so feedback can be provided to assist students in their success, and in the meantime, student advisors can also log into Course Signals to check on the status of their advisees once a signal has been run and to send email to associated students within Course Signals (Purdue University, 2013). Students do not directly use Course Signals but can view their signal (Purdue University, 2013). Researchers' interests in learning analytics are also high to employ research techniques empowered by big data to enhance educational research. Abdous, He, and Yen (2012) conducted a study using data mining for predicting relationships between online question theme and final grade in a live-video integrated college courses. Hung, Hsu, and Rice (2012) integrated data mining in program evaluation of K-12 online education using clustering analysis and decision tree analysis. This study demonstrated "how data mining can be incorporated into program evaluation in order to generate in-depth information for decision making" (p. 27) in addition to specifying the characteristics of successful and at-risk students. These leading practices show that big data can be used at an institutional level as well as a course level.

K-12 schools also adopted data-driven innovations such as predictive analytics (Ravindranath, 2014), personalized data (Fitzgerald, 2013), and cloud-based databases for student information and grade information (Castro, 2013). At the policy-making level, the Center for Data Innovation (2014) urged the government to "encourage data use in the education sector by helping make data available; one approach could be encouraging states to adopt centralized student information databases, from which administrators could conduct analytics and other trusted authorities could develop new educational technologies". This approach could potentially bring big data usage to a larger audience level.

However, as Castro (2013) observed, several states in the United Stated withdrew from storing students' information into "a single, analytics-friendly database" (p. 1) because of the fear of students' data being leaked and students' privacy being breached. These events show that current privacy-protection practices are not sufficient to earn parents' and schools' trust in allowing student data to be part of the cloud-based dataset.

Such setback can also be found in industry, as revealed by the "Netflix Prize" story. Netflix, the online entertainment provider, released some datasets in 2006 "containing 100 million anonymous movie ratings and challenged the data mining, machine learning and computer science community to develop systems that could beat the accuracy of its recommendation system, Cinematch" (Bennett & Lanning, 2007, p. 1). Regarded by the New York Times as a potential business model for other companies (Lohr, 2009), the Netflix Prize engaged researchers to develop predictive modeling techniques and statistical analysis using its data. However, even though the $1 million prize in utilizing big data achieved its success in terms of improving its internal movie recommendation system, Netflix had to cancel this competition in 2010 in response to member privacy concerns regarding Netflix members' movie watching behavior (Lohr, 2010).

Federal Trade Commission (FTC) investigated the Netflix Prize process in terms of violating member privacy. Similar regulations and investigations are also carried out by other government organizations. For example, the federal government provides guidelines on confidentiality and linkage in the National Center for Educational Statistics (NCES) Statistical Standards (U.S Department of Education, Institute of Education Sciences [IES], 2012). Internationally, new regulations on big data usage and analysis are also being considered. Rubinstein (2013) discussed a recent European policy trend regarding big data usage, replacing the European Union's Data Protective Directive. The new regulation, namely the General Data Protection Regulation, "creates new individual rights and imposes new accountability measures on organizations that collect or process data" (p. 74). The Organization for Economic Cooperation and Development (OECD) published Guidelines on the Protection and Transborder Flows of Personal Data in 1980 (Cate & Mayer-Schönberger, 2013), and it is currently reviewing it regulations regarding the use of big data.

Particularly in the education sector, the U.S. Department of Education (2014) updated the Family Educational Rights and Privacy Acts (FERPA) and Protection of Pupil Rights Amendment (PPRA) regulations regarding using online education services, and it established the Privacy Technical Assistance Center (PTAC) to help school administrators, teachers, and parents learn about "data privacy, confidentiality, and security practices related to students-level longitudinal data systems and other uses of student data" (p. 1). The U.S. Department of Education also published best practices for educational stakeholders to protect students' privacy in using online educational services, including software, apps, and web-based tools. Exemplar best practices given by the U.S. Department of Education are "maintain awareness of other relevant federal, state, tribal, or local laws" (p. 7), "be aware of which online educational services [that] are currently being used in your district" (p. 8), and "be transparent with parents and students" (p. 11). FERPA, PPRA, and some other best practices help school administrators, teachers, parents, and students to be protected when using online educational services. Educational researchers and educational analysts should also pay attention to such government regulations.

Other than government organizations, private sectors also initiated discussions on big data and cloud-computing regulations. Microsoft, in 2012, invited policy-makers, industry leaders, and academic researchers in a discussion of data protection, information privacy, and valuable data flows (Cate & Mayer-Schönberger, 2013). The joint effort made by personnel from the government, industries, and academia is a signal that regulating big data privacy protection is a complex and multi-level task.

Another aspect of regulating big data usage is to establish compliance with professional codes of conduct. A number of computational professional organizations have established codes of conduct or codes of ethics (Computer Ethics Institute, n.d.; Information Systems Security Association [ISSA], 2012). While codes of ethics have a positive impact on judgment of data use, current privacy protection, which focuses on personal identifying information, is not enough in the big data era. Secondary uses of big data sets can reverse engineer past, present, and even future breaches of privacy (Richards & King, 2014). Therefore, developing standards of ethical conduct within the big data arena will be essential, as advocated by Crawford, Gary, and Miltner (2014):

> Scientific research that involves drawing on what is euphemistically known as 'passively collected' big data must face difficult questions and develop new ethical frameworks. This is particularly urgent given the leading professional bodies for computing and engineering, the ACM and IEEE, both have ethical guidelines that are almost two decades old. (p. 1666)

It is imperative for educational researchers to understand the opportunities and challenges of using big data in educational context, as well as the regulations, policies, and code of conducts regarding big data collections and big data processes. In the following section, the discussion is focused on big data ethics specifically in educational settings with different stakeholders.

4. Stakeholders

Stakeholders in big data can be generically divided into three types: big data collectors, big data utilizers, and big data generators (Zwitter, 2014). Big data collectors determine and govern the collection, storage, and expiration of data (Zwitter, 2014). Big data utilizers define the purpose of data usage and big data generators are person/natural processes that generate data voluntarily or involuntarily (Zwitter, 2014). Translated to terms in education, big data collectors can be universities and private companies' online education service providers. Big data utilizers include education researchers and decision maker, while big data generators are teachers and students from whom education data is being generated. In the big data era, the power of controlling others' information is highly concentrated in the hands of data collectors and data utilizers. Any leakage or mishandling of information will affect many end users of the Internet.

Dennen (2015) discussed the concerns and privacy that's related to online courses.

She suggested that students and teachers should be involved in the conversation of establishing privacy statements regarding online course information. The authors of this paper echo Dennen's framework in including different stakeholders in establishing privacy agreement. From our discussion, there are some insights that might help educators to consider ethical issues from individual responsibility and power-distribution aspects.

Long and Siemens (2011) differentiated learning analytics and academic analytics when they suggested that academic analytics is at the decision-making level while learning analytics focuses more on the individual learning process and performance. They further broke down learning analytics and academic analytics into different levels, and they summarized the object of analysis that would benefit different stakeholders at each level. For example, learners and faculty would benefit from learning analytics at 1) the course level, including learning analytics such as social network analysis, conceptual development, discourse analytics, and intelligent curriculum, and 2) at the departmental level, when adopting learning analytics, such as predictive modeling and patterns of success/failure. On the other hand, academic analytics at the institutional level, regional (state/provincial) level, and national and international level would benefit administrators, funders, national governments, and education authorities, respectively.

In application, big data mines increasingly larger data sets for important predictions and often surprising insights (Richards & King, 2014). It presents an amazing possibility to usher in a new age of discovery and innovation for mankind (Richards & King, 2014). It was shown in key studies (Whitman & Mattord, 2012) that the major factor of ethical perceptions in decision-making is education, thus it is vital to cultivate the ethical awareness towards a healthy and balanced use of big data and learning analytics in education. In the next section, the discussion focuses on utilizing practical techniques to achieve privacy preservation in education data so that further recommendations can be made to different stakeholders in utilizing big data analytics in education.

5. Privacy preserving data mining with educational data

The big data, if analyzed appropriately and thoroughly, can reveal knowledge from the data. And the newly discovered knowledge helps us make critical decisions (Liu, 2014). Data mining is one of the technologies used to discover knowledge from big data. On the other hand, some of the data contains sensitive and private information, which, if disclosed, may cause ethical and/or privacy concerns. Many parents or school districts have already raised concerns regarding the risks of storing and accessing students' data on cloud-based devices and platforms. Therefore, when analyzing big data, privacy issues should be taken into consideration (Malik, Ghazi & Ali, 2012; Taneja, Khanna, & Tilwalia, 2014). In this section, several key techniques are discussed regarding protecting privacy while processing educational big data.

Various methods have been proposed to protect privacy, such as removing sensitive values (attribute removal) and hiding sensitive values (data hiding). Matzner (2013) suggested using data minimization techniques in terms of the data quality to preserve user privacy. Data minimization means limiting the data collected and retained, and disposing of the data once it is no longer needed. However, these methods lead to information loss. In order to analyze big data while preserving information, a new field, Privacy Preserving Data Mining (PPDM), has been an active research area to address various privacy issues (Panackal & Pillai, 2013). Although it is still debatable whether PPDM approaches address all privacy issues, different PPDM techniques were developed to find efficient protocols to balance privacy, data utility (the use of data), and computational feasibility (easy to compute). These techniques are classified into three categories: perturbation, cryptographic, and anonymization (Panackal & Pillai, 2013).

5.1 Data Perturbation

Data Perturbation is one of the data distortion approaches for privacy protection, which does so by introducing noise from a known distribution. One example is to add certain attributes to a random number with a standard normal distribution. In such cases, the mean of the attribute is preserved while each record differs from the original record in a way that is not recoverable. The data perturbation approach seeks to accomplish masking of individual confidential data elements while making sure that the random noise (e.g., random numbers) preserves the underlying property from the data so that the patterns can still be accurately estimated. These techniques modify actual data values to 'hide' specific confidential data by "adding random noise to confidential, numerical attributes, thereby protecting the original data" (Taneja et al., 2014, p. 1551). The main purpose of data perturbation techniques is to allow legitimate users to access overall statistics (such as average and the Pearson Correlation constant) of big data sets while hiding individual identity from the user (Wilson & Rosen, 2003). For instance, in a simplified case of student records data, a legitimate system user may not be able to access a particular student's grade, but that same user could determine the average grade of the students in one class.

5.2 Cryptographic Techniques

In cryptographic techniques, sensitive data is encrypted. Cryptographic techniques are extensively studied in distributed environments in which the data is distributed across multiple sites (Lipmaa, 2007). Using encryption techniques, data was first encrypted, then transmitted to another end and decrypted (Panackal & Pillai, 2013; Taneja et al., 2014). One approach is pseudonymization, which breaks the link between personal and student record information, giving one or more identifiers to cover the original identity. Encryption is a well-established technique for building pseudonyms (Panackal & Pillai, 2013). It provides a form of traceable anonymity of student health and student discipline records. Instead of completely removing personal identification information from student health and student discipline data, identification information is transferred into a piece of information (i.e., a pseudonym) that cannot be mapped to a student without knowing a certain secret.

5.3 Anonymization

Sometimes it is required that the data must be publically published in its original form without any encryption or perturbation (Panackal & Pillai, 2013). In these cases, before releasing the data, the data needs to be anonymized to protect identity disclosure. There are several methods for obtaining anonymity such as generalization, suppression, data removal, permutation, and swapping (Panackal & Pillai, 2013). One of the common methods is to generalize identifiers. For example, the date of birth of students can be generalized to month of birth. One of the anonymizations is a blocking-based techniques. In blocking-based techniques, it is assumed that there is a sensitive classification rule used for hiding sensitive data from others. In blocking-based techniques, two steps are generally used for privacy protection.

Sensitive data are identified and then replaced by unknown values (?) (Parmar, Rao, & Patel, 2011). For example, actual values can be hidden by replacing ‘1’ by ‘0’ or ‘0’ by ‘1’ or with any unknown (?) values. This provides protection of sensitive data from unauthorized access (Taneja et al., 2014). In student record data, instead of identifying students with social security numbers, using student IDs would increase the level of protection of students’ other records.

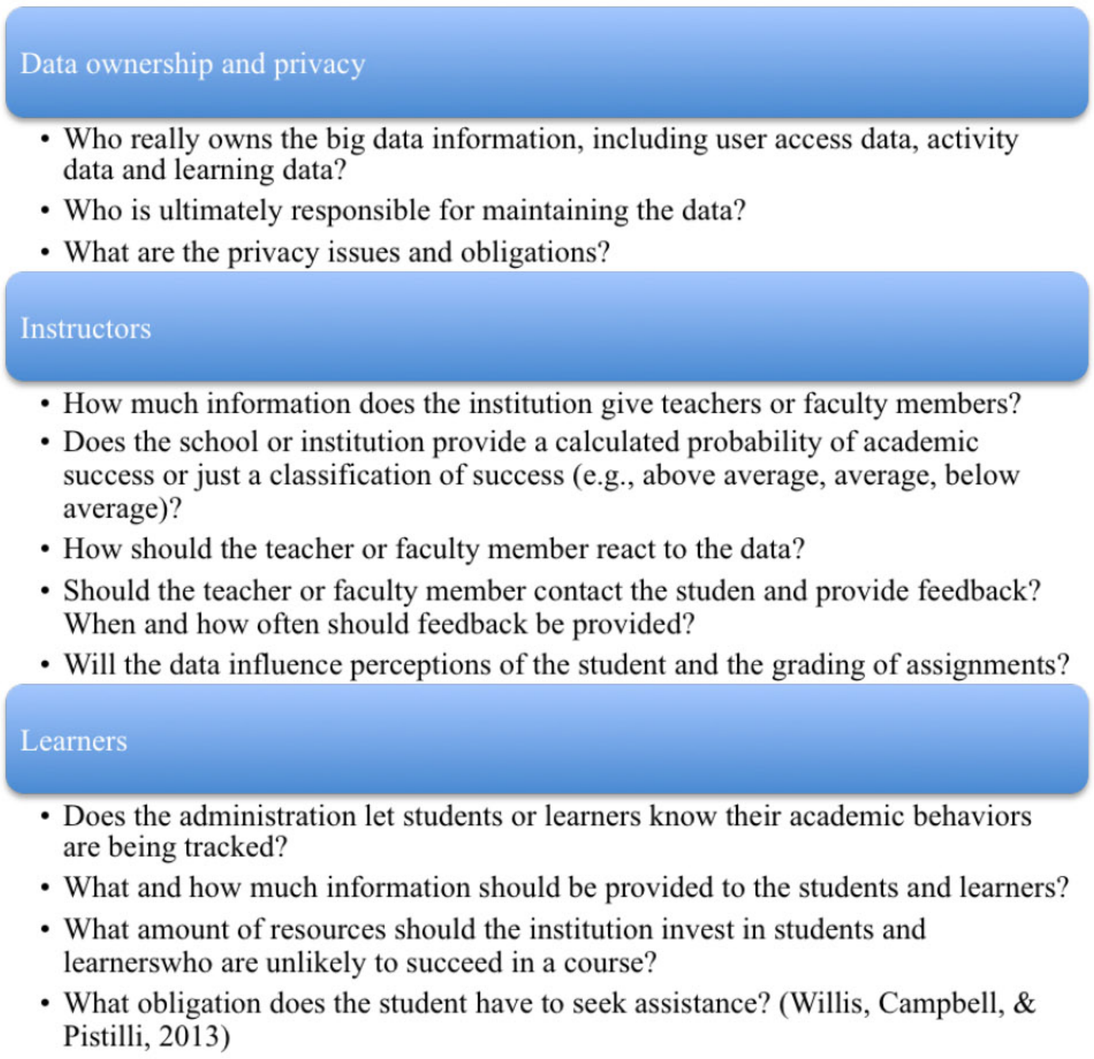
5.4 Summary and Recommendations

The authors argue that in education context, not one single technique might solve all privacy issues, so a hybrid approach is recommended for researchers who are interested in researching big data privacy in future educational data. Each of the approaches for PPDM has its advantages and disadvantages. In perturbation approaches, different attributes are preserved independently. However, the original data values may not be regenerated, which leads to loss of information. In cryptographic approaches, sensitive data is encrypted and transformed, which provides good privacy protection. However, it requires a considerable amount of computation power and therefore is time consuming. In anonymization approaches, sensitive data is hidden, which may also cause loss of information. Therefore, in order to better mine data while preserving privacy, hybrid techniques that combine various approaches such as data perturbation, cryptographic techniques, and anonymization may be developed (Taneja et al., 2014).

6. Discussion

As big data technology progresses, it becomes possible to track every aspect of one’s online life. Transparency is then a characteristic of the current digital world. The collection, processing, and retention of data for analytical purposes have become commonplace in modern business, and consequently the associated legal considerations and ethical implications have also grown in importance (Willis, Campbell, & Pistilli, 2013). Big data ethics are principles that should be recognized as governing data flows in the information society, and should inform the establishment of legal and ethical big data norms (Richards & King, 2014). From its origin in cyberethercs, big data ethics inherits the ethical challenges of a cyber world, and at the same time, the four characteristics of big data add another layer of consideration to ethically handle education data in this age of transparency. The authors of this paper recommend researchers consider the questions shown in Figure 2, from data ownership and privacy, and instructor and learner responsibilities, in applying big data analytics in education, and at the same time, analyst could consider using hybrid approaches of PPDM techniques to enhance the privacy protection at different levels. PPDM considerations are highlighted in Figure 3.

At the privacy and data ownership level, the key ethical consideration would be the power of ownership. Questions regarding who really owns the big data information, including user access data, activity data, and learning data should be addressed before or through the data collection and data-analysis process. Establishing user agreements with different stakeholders in collecting, analyzing, and visualizing data would enhance the partnership in applying data analytics in education. Questions regarding data flow can also be addressed, such as who is ultimately responsible for maintaining the data. At this level, various stakeholders can be involved in the dialogue and process of establishing the agreement, such as learners (young and



Data ownership and privacy

- Who really owns the big data information, including user access data, activity data and learning data?
- Who is ultimately responsible for maintaining the data?
- What are the privacy issues and obligations?

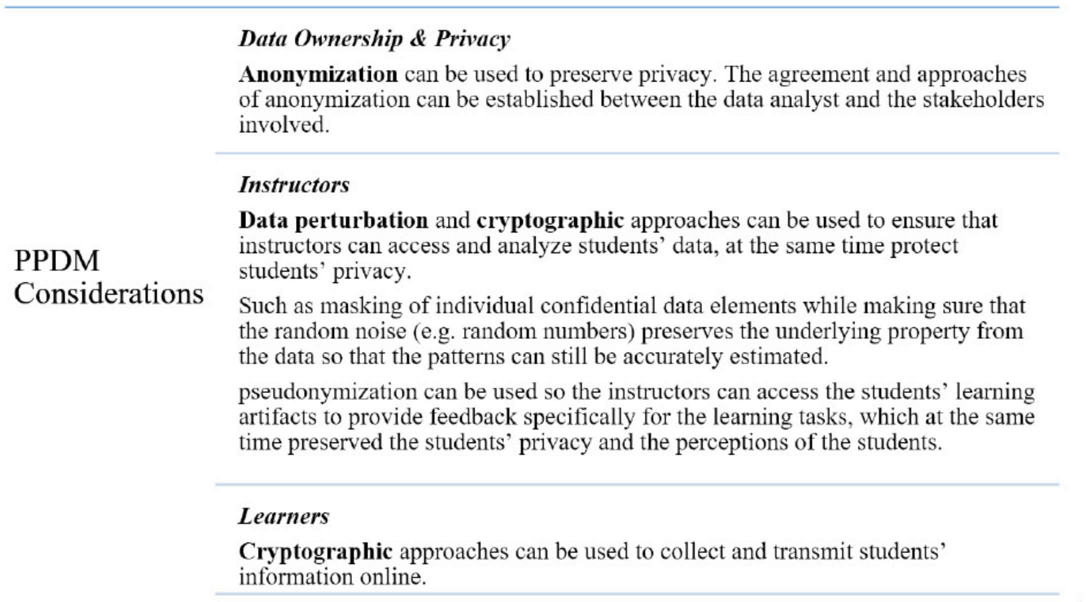
Instructors

- How much information does the institution give teachers or faculty members?
- Does the school or institution provide a calculated probability of academic success or just a classification of success (e.g., above average, average, below average)?
- How should the teacher or faculty member react to the data?
- Should the teacher or faculty member contact the studen and provide feedback? When and how often should feedback be provided?
- Will the data influence perceptions of the student and the grading of assignments?

Learners

- Does the administration let students or learners know their academic behaviors are being tracked?
- What and how much information should be provided to the students and learners?
- What amount of resources should the institution invest in students and learnerswho are unlikely to succeed in a course?
- What obligation does the student have to seek assistance? (Willis, Campbell, & Pistilli, 2013)

Figure 2. Big Data Ethical Consideration in Education.



PPDM Considerations

Data Ownership & Privacy

Anonymization can be used to preserve privacy. The agreement and approaches of anonymization can be established between the data analyst and the stakeholders involved.

Instructors

Data perturbation and **cryptographic** approaches can be used to ensure that instructors can access and analyze students' data, at the same time protect students' privacy.

Such as masking of individual confidential data elements while making sure that the random noise (e.g. random numbers) preserves the underlying property from the data so that the patterns can still be accurately estimated.

pseudonymization can be used so the instructors can access the students' learning artifacts to provide feedback specifically for the learning tasks, which at the same time preserved the students' privacy and the perceptions of the students.

Learners

Cryptographic approaches can be used to collect and transmit students' information online.

Figure 3. Privacy Preserving Considerations of Big Data Ethics.

adult), instructors, parents, institutions, digital learning providers, data analysts, and/or government agencies. Addressing key ethical considerations of the power of ownership can be viewed as one of the fundamental step in ensuring data analysts and other stakeholders establish agreement on the ethical use of big data analytics in different educational contexts. At this level, it is also important for all stakeholders involved to keep up with the most-updated policy and regulations regarding data ownership, data use, and data representations as established by government agencies, such as the U.S. Department of Education. Technically, using PPDM at such level, anonymization can be used to preserve privacy. In student record data, instead of identifying students with social security numbers, using student IDs would increase the level of protection of students' other records. The agreement and approaches to doing so can be established between the data analyst and the stakeholders involved.

At the instructors' responsibility level, the key ethical consideration would be around what practices could the instructor carry out to address possible ethical dilemmas? At this level, ethical considerations could move from the ownership and agreement to the teaching and the use of the data, such as agreement on how much information should the instructors be given, strategies to providing feedback, data influence in grading and perceptions of the students. At this level, the instructors and other data analysts involved would benefit from keeping up with the best practices in the field as established by government agencies or researchers. It is inevitably a life-long learning process. Technically, at the PPDM level, data perturbation and cryptographic approaches can be used to ensure that instructors can access and analyze students' data while protecting students' privacy, such as masking of individual confidential data elements while making sure that the random noise (e.g., random numbers) preserves the underlying property from the data so that the patterns can still be accurately estimated. In other words, data analysts can provide an overall analysis to the instructor, so the learning patterns of the whole class are being captured and analyzed while no specific student information is given out. In addition, pseudonymization can be used so the instructors can access the students' learning artifacts to provide feedback specifically for the learning tasks, which at the same time preserved the students' privacy and the perceptions of the students.

At the student responsibility level, the key ethical concern might be what consequences or strategies would the learners have to carry out on their own? Questions such as does the administration let students or learners know their academic behaviors are being tracked, what and how much information should be provided to the students and learners, and what amount of resources should the institution invest in students and learners who are unlikely to succeed in a course. Technically, at the PPDM level, cryptographic approaches can be used to collect and transmit students' information online. Also educating students' about their responsibilities and best practices in doing so would enhance students' learning.

The discussions and questions at different levels of the hybrid approach of PPDM would add to the existing literature in exploring big data ethics in the educational sphere with societal impact. It is a complex and multi-layer task for all stakeholders involved to work together and address the ethical issues, especially privacy issues in an educational context. The discussion of this paper is to initiate the dialogue of promoting ethical awareness among educators, researchers, and analysts to strive toward a more exciting and protected environment for all learners.

7. Conclusion and Future Research

The rise of big data analytics demonstrates a powerful and democratic vision for education, but it also calls educators, policy makers, administrators, parents, and community partners to assist our learners, children, adolescents, and grown-ups. The increased amount of data and the speed that the data is accumulated provides new challenges to its uses and applications, particularly from the privacy perspective. This paper aims to initiate a dialogue with trends, issues, and concerns regarding employing big data in an educational context. With the discussion on big data ethics, dilemmas with data privacy, policies, cases, and codes of conduct, as well as key privacy-preserving data mining techniques, it is the authors' hope that ethical awareness of using big data can be cultivated at different levels of users, i.e. students, instructors, and the institutions. Future research studies focusing on best practices and guidelines to cultivate big data ethics at individual level and societal level would be valuable in ensuring the ethical use of big data.

References

Abdous, M., He, W., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. Journal of Educational Technology & Society, 15(3), 77–88.

Allen, I. E., & Seaman, J. (2013). Changing course: Ten years of tracking online education in the United States. Newburyport, MA: Sloan Consortium. Retrieved from http://sloanconsortium.org/publications/survey/changing_course_2012

Allen, I. E., & Seaman, J. (2015). Grade level - Tracking online education in the United States. Newsburyport, MA: Online Learning Consortium. Retrieved from http://www.onlinelearningsurvey.com/reports/gradelevel.pdf

Azari, R. (Ed.). (2003). Current security management & ethical issues of information technology. Hershey, PA: IRM Press.

Baase, S. (2013). A gift of fire: Social, legal, and ethical issues for computing technology (4th ed.). Upper Saddle River, NJ: Pearson.

Baker, R. S. (2015). Big data and education (2nd ed.). New York, NY: Teachers College, Columbia University.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), Learning analytics: From research to practice (pp. 61–75). New York, NY: Springer.

Bennett, J., & Lanning, S. (2007). The Netflix prize. Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07), San Jose, California, USA, 1–8. Retrieved from https://www.cs.uic.edu/~liub/KDD-cup-2007/NetflixPrize-description.pdf

Bichsel, J. (2012). Analytics in higher education: Benefits, barriers, progress, and recommendations. Louisville, CO: EDUCAUSE Center for Applied Research. Retrieved from http://net.educause.edu/ir/library/pdf/ers1207/ers1207.pdf

Bourne, P. E. (2014). What Big Data means to me. Journal of the American Medical Informatics Association, 21(2), 194–194.

Cate, F. H., & Mayer-Schönberger, V. (2013). Notice and consent in a world of big data. International Data Privacy Law, 3(2), 67–73.

Castro, D. (2013, December 3). Parents and educators should embrace, not fear, student data. Center for Data Innovation. Retrieved from http://www.datainnovation.org/2013/12/parents-and-educators-should-embrace-not-fear-student-data/

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165–1188.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171–209.

Computer Ethics Institute. (n.d.). Ten commandments of computer ethics. Retrieved from http://computerethicsinstitute.org/publications/tencommandments.html

Cox, M., & Ellsworth, D. (1997). Managing big data for scientific visualization. ACM Siggraph, 97, 21–38.

Crawford, K., Gray, M. L., & Miltner, K. (2014). Critiquing big data: Politics, ethics, epistemology. International Journal of Communication, 8, 1663–1672.

Daniel, B. (2015). Big data and analytics in higher education: Opportunities and challenges. British Journal of Educational Technology, 46(5), 904–920. doi:10.1111/bjet.12230

Dennen, V. P. (2015). Technology transience and learner data: Shifting notions of privacy in online learning. Quarterly Review of Distance Education, 16(2), 45–59.

Einav, L., & Levin, J. (2014). Economics in the age of big data. Science, 346(6210), 1243089. doi:10.1126/science.1243089

Eynon, R. (2013). The rise of Big Data: what does it mean for education, technology, and media research? Learning, Media and Technology, 38(3), 237–240. doi:10.1080/17439884.2013.771783

Fitzgerald, M. (2013, July 26). Big data helps guide Colorado's public schools. InformationWeek. Retrieved from http://www.informationweek.com/

Francisco, A. (n.d.). Realizing the opportunity for big data in education. Retrieved from http://www.digitalpromise.org/blog/entry/realizing-the-opportunity-for-big-data-in-education

Hauptman, R. (1988). Ethical challenges in librarianship. Phoenix, AZ: Oryx Press.

Hung, J.-L., Hsu, Y.-C., & Rice, K. (2012). Integrating data mining in program evaluation of K–12 online education. Journal of Educational Technology and Society, 15(3), 27–41.

IBM. (n.d.). The four V's of big data. IBM – Big Data & Analytics Hub. Retrieved from http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Information Systems Security Association [ISSA]. (2012). ISSA code of ethics. Retrieved from http://www.issa.org/?page=CodeofEthics

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. Communications of the ACM, 57(7), 86–94. doi:10.1145/2611567

Johnson, D. G. (1997). Ethics online: Shaping social behavior online takes more than new laws and modified edicts. Communications of the ACM, 40(1), 60–65. doi:10.1145/242857.242875

Johnson, D. G. (2001). Computer Ethics (3rd ed.). Upper Saddle River, NJ: Pearson.

Johnson, D. G. (2009). Computer Ethics (4th ed.). Upper Saddle River, NJ: Pearson.

Kizza J. M. (2014). Computer network security and cyber ethics (4th ed.). Jefferson, NC: McFarland.

Lane, J. E. (Ed.). (2014). Building a smarter university: Big data, innovation, and analytics. Albany, NY: SUNY Press.

Liu, Y. (2014). Big data and predictive business analytics. Journal of Business Forecasting, 33(4), 18–21.

Lohr, S. (2009, September 21). A $1 million research bargain for Netflix, and maybe a model for others. New York Times. Retrieved from http://www.nytimes.com/

Lohr, S. (2010, March 12). Netflix cancels contest after concerns are raised about privacy. New York Times. Retrieved from http://www.nytimes.com/

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), 31–40. Retrieved from https://net.educause.edu/ir/library/pdf/erm1151.pdf

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. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Malik, M. B., Ghazi, M. A., & Ali, B. (2012). Privacy preserving data mining techniques: Current scenario and future prospects. Proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012, Allahabad, India, 27–32. doi:10.1109/ICCCT.2012.15

Maner, W. (1980). Starter kit in computer ethics. Hyde Park, NY: Helvetia Press and the National Information and Resource Center for Teaching Philosophy.

Maner, W. (1996). Unique ethical problems in information technology. Science and Engineering Ethics, 2(2), 137–154.

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. Jama, 309(13), 1351–1352. doi:10.1001/jama.2013.393

Panackal, J. J., & Pillai, A. S. (2013). Privacy preserving data mining: An extensive survey. Proceedings of the International Conference on Multimedia Processing, Communication and Information Technology, MPCIT (ACEEE), 297–304. doi:03.AETS.2013.4.15

Papamitsiou, Z., & Economides, A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Educational Technology & Society, 17(4), 49–64.

Parmar, A. A, Rao, U. P., & Patel, D. R. (2011). Blocking based approach for classification rule hiding to preserve the privacy in database. Proceedings of International Symposium on Computer Science and Society, IEEE 2011, Kota Kinabalu, Malaysia, 323–326. doi:10.1109/ISCCS.2011.103

Payton, T. M., & Claypoole, T.. (2014). Privacy in the age of big data: Recognizing threats, defending your rights, and protecting your family. Lanham, MD: Rowman & Littlefield.

Picciano, A. G. (2012) The evolution of big data and learning analytics in American higher education. Journal of Asynchronous Learning Networks, 16(3), 9–20

Purdue University. (2013). Course signals. Retrieved from http://www.itap.purdue.edu/learning/tools/signals/

Quigley, M. (Ed.). (2007). Encyclopedia of information ethics and security. Hershey, PA: IGI Global.

Ravindranath, M. (2014, January 12). D.C. Prep uses big data to evaluate tablet-based education apps. The Washington Post. Retrieved from https://www.washingtonpost.com/

Richards, N. M., & King, J. H. (2014). Big data ethics. Wake Forest Law Review, 49, 393–432.

Rubinstein, I. S. (2013). Big data: The end of privacy or a new beginning? International Data Privacy Law, 3(2), 74–87. doi:10.1093/idpl/ips036

Schultz, R. A. (2006). Contemporary issues in ethics and information technology. Hershey, PA: IRM Press.

Sejnowski, T. J., Churchland, P. S., & Movshon, J. A. (2014). Putting big data to good use in neuroscience. Nature Neuroscience, 17(11), 1440–1441.

Shepherd, M., & Watters, C. (2013). What does knowledge organization mean in a big data environment? SRELS Journal of Information Management, 50(6), 819–829. doi:10.17821/srels/2013/v50i6/43830

Smolan, R., & Erwitt, J. (2012). The human face of big data. Sausalito, CA: Against All Odds Productions.

Spinello, R. A. (2000). Cyberethics: Morality and law in cyberspace. Boston, MA: Jones and Bartlett.

Spinello, R. A., & Tavani, H. T. (Eds.). (2004). Readings in cyberethics (2nd ed.). Sudbury, MA: Jones & Bartlett.

Taneja, S., Khanna, S., & Tilwalia, S. (2014). A review on privacy preserving data mining: Techniques and research challenges. International Journal of Computer Science and Information Technologies, 5(2), 2310–2315

Tavani, H. T. (2010). Ethics and technology: Controversies, questions, and strategies for ethical computing (3rd ed.). Hoboken, NJ: John Wiley & Sons.

U. S. Department of Education, Institute of Education Sciences. (2012). 2012 Revision of NCES statistical standards. Retrieved from http://nces.ed.gov/statprog/2012/

U. S. Department of Education. (2014). Protecting student privacy while using online educational services: Requirements and best practices (Report No. PTAC-FAQ-3). Retrieved U. S. Department of Education – Privacy Technical Assistance Center website: http://ptac.ed.gov/document/protecting-student-privacy-while-using-online-educational-services

Vaitsis, C., Nilsson, G., & Zary, N. (2014). Big data in medical informatics: improving education through visual analytics. Stud Health Technol Inform, 205, 1163–7.

Wainer, H., Gruvaeus, G., & Blair, M. (1974). TREBIG: A 360/75 FORTRAN program for three-mode factor analysis designed for big data sets. Behavior Research Methods & Instrumentation, 6(1), 53-54.

Whitman, M., & Mattord, H. (2012). Principles of information security (4th ed.). Boston, MA: Cengage Learning.

Wiener, N. (1948). Cybernetics, or control and communication in the animal and the machine. New York, NY: Technology Press/John Wiley & Sons.

Wiener, N. (1950). The human use of human beings: cybernetics and society. Boston, MA: Houghton Mifflin.

Wiener, N. (1964). God & Golem, Inc.: A comment on certain points where cybernetics impinges on religion. Cambridge, MA: MIT Press.

Willis, J. E., Campbell, J. P., & Pistilli, M. D. (2013). Ethics, big data, and analytics: A model for application. EDUCAUSE Review Online. Retrieved from http://www.educause.edu/ero/article/ethics-big-data-and-analytics-model-application

Wilson, R. L., & Rosen, P. A. (2003). Protecting data through 'perturbation' techniques: The impact on knowledge discovery in databases. Journal of Database Management, 14(2), 14–26.

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. Knowledge and Data Engineering, IEEE Transactions on, 26(1), 97–107.

Yang, C., Hung, J.-L., & Lin, Z. (2013). An analysis view on password patterns of Chinese internet users. Nankai Business Review International, 4(1), 66–77.

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. Ethics and information technology, 12(4), 313–325.

Zwitter, A. (2014). Big data ethics. Big Data & Society, 1(2), 1–6. doi:10.1177/2053951714559253

Contact the Author

Xiaojun Chen
Department of Curriculum and Instruction,
School of Education, St. John's University
Email: chenx@stjohns.edu

Ying Liu
Division of Computer Science, Mathematics
and Science, College of Professional Studies,
St. John's
Email: liuy1@stjohns.edu