Privacy in the era of big data and learning analytics: ethical considerations and positions

Marianna Vivitsou
Faculty of Behavioral Sciences,
University of Helsinki
Siltavuorenpenger 5A
FI-00014 Helsinki
+35 8 50 318 5358
marianna.vivitsou@helsinki.fi

Mohsen Saadatmand
Faculty of Behavioral Sciences,
University of Helsinki
Siltavuorenpenger 5A
FI-00014 Helsinki
+358442810707
mohsen.saadatmand@helsinki.fi

ABSTRACT
In the era of big amounts of new data, the need arises to reconsider the role of higher education institutions and institutional policies in relation to privacy, data control and trust. It is then important to gain better insight into the context of this new way of analysis and thus contribute to the process of building an agenda about data ethics and privacy. Considering these, in this paper we are interested in discussing and putting forward our position regarding context integrity in learning analytics. To do so, we will discuss the elements of context integrity in relation to the type of research conducted online and emerging challenges.

CCS Concepts
• Security and privacy~Privacy-preserving protocols
• Security and privacy~Access control
•Security and privacy~Pseudonymity, anonymity and untraceability

Keywords
Context integrity; privacy; big data; learning analytics; institutional policies and practices.

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

1. INTRODUCTION
In the era of massive (or big) amounts of new data, the need arises to reconsider the role of higher education institutions and institutional policies in relation to privacy, data control and trust. Big data is intertwined with analytics aiming for increased effectiveness in learning. It is then important to gain better insight into the context of this new way of analysis and thus contribute to the process of building an agenda about data ethics and privacy. Considering these, in this paper we are interested in discussing and putting forward our position regarding context integrity in learning analytics. To do so, it is necessary to examine the elements that constitute context integrity in relation to the type of research conducted online and the challenges that emerge there.

2. OVERVIEW OF THE FIELD OF INTEREST
In order to contribute to an agenda about data ethics and privacy in the effort of seeking ways for better learning analytics, in this position paper we will discuss privacy in relation to context integrity. According to Ohm [10], integrity is the social context where research is conducted and where the challenges to privacy come up. In other words, as Nissenbaum [9] suggests, context integrity is shaped where privacy is shaped. Challenges to privacy jeopardize institutional trust in the sense that commitment to a code of moral values is disrupted. It is essential, therefore, to understand what constitutes moral practice in learning analytics [12]. Slade and Prinsloo [12] argue that in higher education where structural design allows a central authority to oversee the totality of institutional activity, it is administration and faculty decision making and action that is more likely to compromise trust than the action of students by, for instance, mining data from institutionally managed software aiming to provide insight into student performance.

One aspect of context integrity, therefore, relates to underlying practices and power relations when the purpose is, for instance, to activate the predictive power of analytics by making use of systems that enable multi-directional surveillance and target student performance data [4, 12]. Under this lens, trust increase or decrease happens on the basis of institutional initiative and that makes very little, if at all, room for other parties to take action and exert control. Trust, however, should not be treated as a unidirectional property [5]. As the established notion of privacy

collapses, it is necessary to view trust not as stable entity. Trust is a dynamic process that involves the perspectives of different actors, is amenable to negotiation on an ongoing basis, and considers both the technological system and infrastructure, and the policies that regulate institutional life as social domain [6]. In many respects, dynamic views seem to fit well with the complexities of the current situation.

Nowadays, as universities expand their Internet-connected services and the ways these services are offered, the situation becomes even more complex. On the one hand, it is now possible for students to access their library accounts, gym bookings, or university provided social media services using any possible mobile, or not, device. On the other hand, it is also possible for researchers to opt Internet-connected devices as data collection points, without the consent of the users of those devices. In addition to students, faculty and administration also leave digital traces [4], thus shifting information transmission set practices and making the communication of systems (i.e., institutions and ICTs) (or their interoperability) even more challenging at technical, semantic, legal, and political levels [5].

As informational norms get disrupted, more considerations for privacy and systems design arise. In their case study on a censorship measurement code, Narayanan & Zevenbergen [8] argue that researchers can alter the behavior of Internet-connected devices in order to gain scientific data about the behavior of users and networks. University virtual spaces and online services are often quite effectively firewalled. Nowadays, however, we are aware that firewall protection does not suffice to avoid all types of intrusiveness.

Disruptions in information flows generate skepticism as to the best ways to deal with the new situation and what social values are served or threatened [8]. In this way, a wider issue concerning research ethics in learning analytics opens up. It seems, therefore, that along with transmission practices, research ethics in higher education must change [4, 12] as well. One way to understand the directions of this change is by reviewing the existing landscape and re-examine needs in order to adapt policies and structures in such ways that higher education institutions can best meet the requirements of the big data era.

To this end, in the following section we will discuss issues and questions arising in the current situation in relation to existing practices in research ethics, the impact of changes on anonymity and consent, as well as validation mechanisms in computer science-based research.

3. ISSUES THAT NEED TO BE ADDRESSED

3.1 Issue 1: Existing Practices

So far, ethical principles in research have been handled by Advisory Boards (ABs) issuing official documents that define responsible conduct of research. In Finland, for example, such documents state that privacy and data protection should intertwine with the goal to reach a balance between confidentiality and the openness of science and research [3].

On the one hand, AB specifications, as apply to the Finnish context, nowadays seem to be outdated. Typically, ABs approve research proposals to account for and mitigate risk to participants. However, it seems that computer science-based research efforts have been exempted from extensive oversight in, for instance, the U.S. [8]. As Narayanan & Zevenbergen [8] argue, university ABs are geared toward regulating mainly biomedical and social sciences research. Our review of the system that regulates the Finnish context [2] is in agreement with this argument.

Certainly, ABs do well in overseeing the previously mentioned fields of research. The way things are scaling up nowadays, however, makes this kind of oversight inadequate. Existing specifications, for example, relate to the aspect of the university as science and research organization that allows for data collection and databases creation toward knowledge growth and advancement.

What happens when the whole university becomes the data? What would an agenda for ethics and privacy concerning mining data that involve researchers' students' and teachers' profiles, views and behaviors look like?

3.2 Issue II: Anonymity and Consent

Evidently, what is played out here is the right of the user (i.e., student, researcher or teacher) to remain anonymous and choose whether she grants permission for data use for research purposes. Anyway, this is how things have been handled to date in order to secure 'responsible conduct of research'. Anonymity and informed user consent guarantee respect for the research subjects' rights and subject control over own data. Two issues arise here.

One concerns how the notion of anonymity changes in the new circumstances. Findings of relevant studies [1] have shown that sensitive user facts can be statistically inferred and, therefore, data anonymization is simply not enough.

The other issue concerns informed consent. Barocas & Nissenbaum [1] argue that the way things turn out (e.g., challenges in operationalizing big data; unpredictability as to fruitful results, for instance, whether analysis leads to intuitive correlations or not), the need for consent becomes questionable. To add to this, we should mention here Prinsloo and Slade's view [11] that, since consent is under-theorized, it is the context and user understanding of context that determine the scope and detail of information sharing. User agency and privacy self-management [11] are, therefore, critical.

Considering privacy issues in relation to the way the situation evolves, these questions arise:

What are the criteria for determining proper and improper uses of big data collection and analysis? Who should oversee the integration of such standards?

3.3 Issue III: Validation of Research

In practice, validation of research in ethical grounds is performed by conference committees and entails for-publication-studies. In these cases, computer security, networking and Internet measurement sub-communities converge on conference program committees and act as oversight mechanisms. The system seems to undergo several shortcomings. One is that this kind of ethical review is retrospective and, as such, fails to prevent harm arising from conducting, rather than publishing, the research [8]. Another is that big data collection and analysis is not exclusively the business of computer scientists. Massive data analytics can impact different aspects of human subject lives and, therefore, should integrate insights from the social sciences and the humanities. However, program committees rarely include members with scholarly expertise in research ethics and ethics in general.

One reason to explain the situation relates to the indeterminacy to arrive at a definition of computer science-based and engineering

research. Is this research targeting human or human-less systems? What ethical concerns should be involved? And what mechanisms should validate research conduct?

4. DISCUSSION

Overall, based on our review of the background literature in the previous section, we understand that the prevailing view sees context integrity as a set of processes that extend along the continuum of data collection, data control, and analysis and report. As Figure 1 shows, when dealing with big data, we need to think of certain procedures (e.g., ethical guidelines) and how to deal with them. Procedures and techniques ensure informed consent and ethical guidelines (whether set by the university, advisory board or researcher) should be taken care of throughout the data collection, analysis and reporting processes in the continuum. Concerns for privacy, ownership and anonymity also appear throughout the different parts.

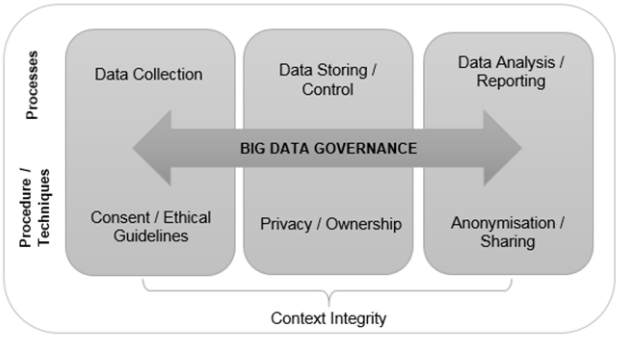

Figure 1. A representation of context integrity in learning analytics

Questions posed in the continuum are dynamic and the responses sought should respect the human subject autonomy and consider her right to choose on an ongoing basis. As this aspect of context integrity focuses on the conditions underlying production and disclosure of data, it mainly raises issues of transparency. Transparency is indeed a core value in institutional governance [4], but it is only half way through a holistic approach to context integrity.

We believe that a consideration is needed in terms of the opportunity opened up for students, faculty and administration to develop a broad perspective and responsibility to social welfare. Toward this end, in this position paper we put forward the need to review the current situation in research ethics and ethics education, and re-examine training programs in order to adapt curricula in such ways that the requirements of the big data era can be successfully met.

In the core of such position lies the view of learning analytics as a set of moral practices and the supposition that what is needed is an understanding of the scope, the role and the boundaries of learning analytics. To get there, as Slade and Prinsloo [12] suggest, we need to regard both what is effective and what is appropriate by establishing cultures of both measuring and understanding. In this way, integrity will be served not only as adherence to sets of moral codes but also in terms of soundness and completeness when the overall institutional role and mission come into play.

Toward this end, we propose the following:

- The role of Advisory Boards should be updated and expanded in order to fit current needs arising from the developments in big data and learning analytics.
- Computer science-based research is research targeting humans and both technological system and research design should consider this principle.
- Massive data analytics should integrate insights from the social sciences and the humanities and all interested parties should be involved in the design and decision making processes.
- University pedagogies should draw links with responsible conduct of research with regard to current needs. Therefore, university curricula should enable scholarly expertise in research ethics and ethics in general in relation to conducting internet-based research. In this way, pedagogical approaches in data ethics can enable higher education stakeholders to develop a broad perspective and responsibility to social welfare.

5. REFERENCES

[1] Barocas, S. & Nissenbaum, H. (2014). Big Data's End Run around Anonymity and Consent. In Lane, J., Stodden, V., Bender, S., & Nissenbaum. H. (Eds.). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Cambridge University Press: NY.

[2] Finnish Advisory Board on Research Integrity (2012). *Responsible conduct of research and procedures for handling allegations of misconduct in Finland*. Accessed: February 11, 2016. http://www.tenk.fi/sites/tenk.fi/files/HTK_ohje_2012.pdf

[3] Finnish Advisory Board on Research Ethics (2009). Ethical principles of research in the humanities and social and behavioral sciences and proposals for ethical review. Accessed: February 11, 2016. http://www.tenk.fi/sites/tenk.fi/files/ethicalprinciples.pdf

[4] Knox, D. (2010). A good horse runs at the shadow of the whip: Surveillance and organizational trust in online learning environments. Canadian Journal of Media Studies. Accessed: February 15, 2016. http://cjms.fims.uwo.ca/issues/07-01/dKnoxAGoodHorseFinal.pdf

[5] Hoel, T. and Chen, W. (2014) Learning Analytics Interoperability – looking for Low-Hanging Fruits in Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education. Presented at The 1st ICCE Workshop on Learning Analytics (LA2014), Nara, Japan, December 2014.

[6] Hoel, T. and Chen, W. (2015) Privacy-driven design of Learning Analytics applications – exploring the design space of solutions for data sharing and interoperability LAK'15, March 16–20, 2015, Poughkeepsie, NY, USA

[7] Metcalf, J., Crawford, K. & Keller, E. F. (2016). "Pedagogical Approaches to Data Ethics." *Council for Big Data, Ethics, and Society*. Accessed February 11, 2016. http://bdes.datasociety.net/council-output/pedagogical-approaches-to-data-ethics-2/.

[8] Narayanan, A. & Zevenbergen, B. (2016). "Case Study: No Encore for Encore? Ethical questions for web-based

censorship measurement."*Council for Big Data, Ethics, and Society*. Accessed: February 11, 2016. http://bdes.datasociety.net/council-output/case-study-no-encore-for-encore/.

[9] Nissenbaum, H. (2010). Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford, CA: Stanford Law Books.

[10] Ohm, P. (2014). Changing the Rules: General Principles for Data Use and Analysis. In Lane, J., Stodden, V., Bender, S., & Nissenbaum. H. (Eds.). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Cambridge University Press: NY.

[11] Prinsloo, P. & Slade, S. (2015). Student vulnerability, agency & learning analytics: an exploration. Presentation at the LAK15 Workshop (EP4LA), March 16, 2015. Poughkeepsie, NY, USA. Accessed: February 11, 2016. http://www.slideshare.net/prinsp/lak15-workshop-vulnerability-final

[12] Slade, S., & Prinsloo, P. (2013). Learning Analytics: Ethical Issues and Dilemmas. American Behavioral Scientist, 57(10), 1510–1529. doi:10.1177/0002764213479366