Researching Big Data Research: Ethical Implications for IS Scholars

Emerging Research Forum Paper

Marco Marabelli
Bentley University, 175 Forest St.
Waltham, MA, 02452, USA
mmarabelli@bentley.edu

M. Lynne Markus
Bentley University, 175 Forest St.
Waltham, MA, 02452, USA
mlmarkus@bentley.edu

Abstract

This ERF (Emerging Research Forum) paper focuses on the ethical implications of IS academic big data research. We explore how big data research raises concerns about privacy, human subjects protection and research integrity that are not yet adequately addressed by law, regulation, or the norms of acceptable research conduct. The objective is to increase awareness and promote constructive debate, with the ultimate goal of developing consensus in the field about appropriate research data use practices.

Keywords: Big data; ethical issues; data fabrication; re-identification; discrimination; privacy; research replication

Introduction

The last decade has witnessed the widespread diffusion of new digital technologies capable of recording the minutiae of individuals' everyday lives (Hedman et al. 2013): "Global IT has enabled information on most everything to flow most everywhere at stealth speed" (Nolan 2012, p. 91). In addition, the increased potential to process and store information has enabled what some call automated or algorithmic decision-making processes (Newell and Marabelli 2015). These processes have been described as "computerized knowledge-based processes that can operate on their own with minimal human intervention" (Markus 2015, p. 58).

Similar to most other technologies, big data analytics and algorithmic decision-making have both a bright and a dark side (de Sola Pool 1983; Markus 2014). On the positive side, these technologies enable businesses to profile customers, identify trends and make quick and effective (and almost automatic) marketing decisions. However, they also raise concerns about privacy threats, control losses (Bélanger and Crossler 2011; Coll 2014) and discrimination (Yoo 2010).

We are not the first scholars to discuss the risks, as well as the advantages, of utilizing big data analytics in academic research (e.g., Barocas and Nissenbaum 2014; King 2015; Metcalf 2016). The risks involved are compounded by the "publicness" of some big data (e.g., social media postings), combined with the proprietary ownership of big data platforms, which provide opportunities for: privacy harms to human research subjects (Lease et al. 2013; Narayanan and Shmatikov 2008), non-consensual experiments (Grimmelmann 2015; Jouhki et al. 2016; Kleinsman and Buckley 2015; Shaw 2016) and the possible shielding of academic misconduct (data falsification/fabrication) (Mattioli 2014). At the same time, we believe that the IS field has not sufficiently explored the ethical issues raised by big data research and what we as a community should do in response to the risks.

The main question we address in this paper is: *What is the emerging consensus in business research about ethical issues related to collection, analysis and potential manipulation of large datasets involving human subjects?* We introduce three interrelated situations in which big data research raises ethical concerns and show that consensus about how to address the concerns is low. We conclude with a discussion of implications for the IS community.

Re-identification Research

De-identification refers to policies and processes aimed at preventing a person's identity from being (re) associated with information. De-identification is common practice in academic research involving human subjects, and it is used to ensure study participants' privacy. Here, we focus on the reverse procedure of *re*-identification, in which researchers aim to re-link data with their owners. Numerous re-identification research studies have been published. For example, successful re-identification "tests" were undertaken against the "Netflix Prize" dataset which contained the anonymous movie ratings of over 500,000 Netflix subscribers (Narayanan and Shmatikov 2008). Similarly, triangulation techniques that exploit second degree connections in social media have shown that it is possible to re-identify users' posts over time (Hay et al. 2008).

Opportunities for intended or inadvertent re-identification during the course of big data research are legion. "Digital traces" give researchers the means to surveil people's actions. For instance, geo-tagging (e.g., Facebook, Instagram) allows an individual's location to be pinpointed at a certain moment in time (Girardin et al. 2008), and these digital traces are "here to stay" (Marabelli et al. 2016). Data published on social media (Hedman et al. 2013; Venturini and Latour 2010) and data from website use and surveillance cameras are increasingly analyzed by by governments and private companies with the collaboration of academic researchers. Amazon does not ensure the anonymity of crowdworkers on its Mechanical Turk platform (https://www.mturk.com/) (Lease et al. 2013), and there are no specific policies prohibiting researchers who use the platform from re-identifying crowdworkers (Benitez and Malin 2010).

There is little consensus in the academic community about the degree to which the possibility of re-identification poses risks to human subjects or whether deliberate re-identification research is ethical. Some claim that anonymized data and "public" or "secondary" social media data are fair game for researchers, and that the ethical responsibilities of researchers are minimal if the data were collected by a third party as in the case of weblogs/digital traces, e.g., on website visits and purchases but also social media feeds (Grimmelmann 2015).

Others disagree. They argue that data in private databases should be treated as primary data and subject to institutional review board (IRB) protections even when the data are not publicly available (Grimmelmann 2015). They assert that reviewers must be provided with detailed information on research procedures involving human subjects, even when the subjects are black hat hackers (see the Menlo report, https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf). They propose the use of data mining mechanisms for privacy protection, which should make it difficult to re-identify users on the basis of their uploaded content (e.g., on Facebook and Twitter) (Blum et al. 2013; Dwork et al. 2006).

But calls for an enhanced regime of privacy protection are frustrated by fragmented laws and regulations, which are often country-specific (MacDonald and Streatfeild 2014; Otjacques et al. 2007). (Even within countries these policies and regulations are not consistent across research fields (Asgary and Mitschow 2002; Scholtens and Dam 2007).) As big data research becomes increasingly important in the field of information systems, concerns about re-identification and risks to personal privacy are bound to grow. It's timely now to begin working toward consensus in our field on appropriate research practices and necessary safeguards.

Big Data Experiments

A second situation in which big data research raises ethical concerns is that of experiments conducted on social media and e-commerce platforms. A much discussed example is the Facebook mood experiment, in which users' news feeds were filtered of either positive or negative news to assess the impacts on users' emotional states (Jouhki et al. 2016). This experiment was undertaken with neither IRB approval nor users' consent (or post-hoc debriefing), despite the fact that university-based researchers participated in data analysis. In an analysis of the Facebook mood study, Grimmelmann (2015) concluded that social media experiments afford opportunities for "IRB laundering:" Facebook (and other social media companies) can capitalize on loopholes in the current system of regulating research to circumvent human

subject protections. Grimmelmann's proposed solution is a lightweight IRB framework tailored specifically to social media experiments (2015).

However, critics have noted numerous problems with the IRB process when it comes to big data experiments and other social media research. First, IRB protection does not extend to human *non-subjects*, who might be indirectly affected negatively by big data research. For example, although a subject may consent to provide tagged photographs, other people in those photos may not have provided consent for their names and images to be used. Second, the IRB process has been contested because: 1) rules differ across countries; some countries have no rules or review committees; 2) rules often apply only to federally funded research; 3) rules are not applied consistently across institutions; and 4) definitions of "existing data" and "quality improvement" efforts (vs. research) are fraught. For instance, some authorities argue that users of password-protected social media sites have reasonable expectations of privacy, while others argue that lawmakers eventually will recognize the ownership rights of data providers, so that websites will not be able to sell these data or distribute them without permission.

Indeed, critics of the IRB process have called for major revisions to the Common Rule on human subjects research in the U.S.: A Notice of Proposed Rulemaking (NOPR) proposes the exclusion of much big data research from IRB reviews (https://www.hhs.gov/ohrp/regulations-and-policy/regulations/nprm-home/). (This initiative was promptly contested by authorities such as the Data & Society think tank (https://datasociety.net) and the RAND Corporation (http://www.rand.org).)

Thus, the current context exhibits strong value conflicts among researchers as well as in society at large. In this context, the likelihood is high that some academic big data studies will continue to generate public controversy, as the Facebook mood experiment did. Because such controversies can threaten the reputation of the academy in general and our field in particular, it is timely for us to begin public discussions about acceptable and unacceptable research practices.

Research Integrity

A third situation in which big data research raises ethical concerns is that of research verification and replication. Big data research increasingly makes use of proprietary data bases and analytic resources to which other researchers cannot readily gain access for purposes of review and replication. This raises the possibility that limited access to data and platforms will be used as a shield for intentional academic misconduct such data falsification or data fabrication.

Claims of privileged access to unique data sources have figured in recent research scandals that did not involve big data. For instance, former Bentley University Accounting Professor James E. Hunton was charged with fabricating research data over several decades. To avoid censure by editors and professional associations, he hid behind an alleged confidentiality agreement that prevented him from providing information that could be used to verify the integrity of his research. He claimed that "disclosure of either the data or the identities of the firms would result in him being subject to lawsuits, to the loss of his CPA license, and to a loss of confidence in the field and thus access to further research opportunities" (https://www.bentley.edu/files/Hunton%20report%20July21.pdf). At the time of writing, Hunton has a total of 37 retractions, making him #8 on the Retraction Watch Leaderboard (of most retracted authors) (http://retractionwatch.com/the-retraction-watch-leaderboard/).

Although this sad case did not involve big data, it is clear that big data research offers new and improved opportunities for academic fraud through privileged access to proprietary data sources and analysis platforms that cannot be verified by reviewers or replicated by other scholars. Because of such concerns, PLOS journals (https://www.plos.org/publications) require all their authors to make the data underlying the findings described in their manuscripts fully available without restriction; an "availability statement" must be provided as part of the submission process, and refusal to share data and related metadata methods is grounds for rejection. PLOS journals do not make exceptions for authors who claim that their career interests prevent them from sharing their data.

Data falsification and fabrication represent a significant problem for the research community. Not only do they degrade the quality of the cumulative knowledge base, they also threaten the reputation of the academic community. Trust in the academy is essential for on-going public support of universities and research. In the current political climate, which devalues science, technical expertise and intellectual

attainment, future research scandals centered on big data would threaten our livelihoods and social mission. It is certainly timely for the IS field to engage in discussion of ways to ensure the integrity of our big data research.

Big Data Research Issues: A Challenge for IS Scholars

When scholars discuss big data risks, in addition to big data advantages, we often discuss the actual or potential actions of corporations and their marketing behavior. But the examples mentioned in this paper suggest that we also need to consider the risks of *academic* big data research and the ethical challenges facing academic researchers. It is uncomfortable to turn the spotlight towards ourselves, but we believe it is necessary, as the reputation of the academy is at stake.

As a first step on this journey, we authors have launched an initiative in our own university to discuss the issues laid out in this paper. We have organized a series of three webinars over the next two months, involving colleagues from departments across the university and external experts, in which the issues of re-identification research and experimentation, the legal regime of privacy and human subjects protection, and academic integrity will be discussed. We will report on the findings from these webinars at AMCIS. The longer-term objective of this work is to develop a framework for comparing and evaluating ethical guidelines for big data research.

As a second step, we plan to engage members of the IS field in dialog about our codes of research conduct relative to other fields. The biomedical field has highly developed sensitivity to the ethical, legal and social implications of genetics research, and the healthcare field has highly developed policies around research conflicts of interest (PEW 2013). Similarly, the ethics of computing is a well-established research domain, and some data science codes of conduct address the issues raised in this paper. By contrast, discussion of these issues and the development of relevant guidelines appears to be lacking in various business and management disciplines, including IS. We believe that much can be learned by comparing codes of conduct and considering ways to harmonize them with due attention to field-specific differences in phenomena of interest and methodology.

Much needs to be done, we believe, to ensure that the rights of human subjects are not overlooked in big data research (Metcalf and Crawford 2016) and that the academic community makes appropriate use of the novel data collection and analysis technologies of big data (Ekbia et al. 2015), thereby safeguarding the integrity of the cumulative knowledge base. While we believe that is premature for our journals and the Association for Information Systems journals to propose strong new rules of ethical big data research conduct, it is *not* too early for us to raise awareness about the ethical challenges and to start building consensus about appropriate behavior.

REFERENCES

Asgary, N., and Mitschow, M. C. 2002. "Toward a Model for International Business Ethics," *Journal of Business Ethics* (36:3), pp. 239-246.

Barocas, S., and Nissenbaum, H. 2014. "Big Data's End Run around Procedural Privacy Protections," *Communications of the ACM* (57:11), pp. 31-33.

Bélanger, 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. 1017-1042.

Benitez, K., and Malin, B. 2010. "Evaluating Re-Identification Risks with Respect to the Hipaa Privacy Rule," *Journal of the American Medical Informatics Association* (17:2), pp. 169-177.

Blum, A., Ligett, K., and Roth, A. 2013. "A Learning Theory Approach to Noninteractive Database Privacy," *Journal of the ACM (JACM)* (60:2), p. 12.

Coll, S. 2014. "Power, Knowledge, and the Subjects of Privacy: Understanding Privacy as the Ally of Surveillance," *Information, Communication & Society* (17:10), pp. 1250-1263.

de Sola Pool, I. 1983. *Forecasting the Telephone: A Retrospective Technology Assessment*. Norwood, NJ: Ablex.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis," *Theory of Cryptography Conference*: Springer, pp. 265-284.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., Tsou, A., Weingart, S., and Sugimoto, C. R. 2015. "Big Data, Bigger Dilemmas: A Critical Review," *Journal of the Association for Information Science and Technology* (66:8), pp. 1523-1545.
Girardin, F., Blat, J., Calabrese, F., Dal Fiore, F., and Ratti, C. 2008. "Digital Footprinting: Uncovering Tourists with User-Generated Content," *Pervasive Computing, IEEE 7.4*), pp. 36-43.
Grimmelmann, J. 2015. "The Law and Ethics of Experiments on Social Media Users," *Colorado Technology Law Journal* (13), pp. 219-272.
Hay, M., Miklau, G., Jensen, D., Towsley, D., and Weis, P. 2008. "Resisting Structural Re-Identification in Anonymized Social Networks," *Proceedings of the VLDB Endowment*, pp. 102-114.
Hedman, J., Srinivasan, N., and Lindgren, R. 2013. "Digital Traces of Information Systems: Sociomateriality Made Researchable," *Proceedings of the 34th International Conference on Information Systems*, Milan, Italy.
Jouhki, J., Lauk, E., Penttinen, M., Sormanen, N., and Uskali, T. 2016. "Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics," *Media and Communication* (4:4), pp. 75-85.
King, J. L. 2015. "Humans in Computing: Growing Responsibilities for Researchers," *Communications of the ACM* (58:3), pp. 31-33.
Kleinsman, J., and Buckley, S. 2015. "Facebook Study: A Little Bit Unethical but Worth It?," *Journal of Bioethical Inquiry* (12:2), pp. 179-182.
Lease, M., Hullman, J., Bigham, J. P., Bernstein, M. S., Kim, J., Lasecki, W., Bakhshi, S., Mitra, T., and Miller, R. C. 2013. "Mechanical Turk Is Not Anonymous," (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2228728 last time accessed February 7th 2017).
MacDonald, D. A., and Streatfeild, C. M. 2014. "Personal Data Privacy and the Wto," *Houston Journal of International Laws.* (625), pp. 625-653.
Marabelli, M., Newell, S., and Galliers, R. D. 2016. "The Materiality of Impression Management in Social Media Use: A Focus on Time, Space and Algorithms," *36th International Conference of Information Systems*, Dublin, Ireland: AIS.
Markus, M. L. 2014. "Information Technology and Organizational Structure," in *Information Systems and Information Technology, Computing Handbook, Volume Ii,* H. Topi and A.B. Tucker (eds.). Chapman and Hall, CRC Press.
Markus, M. L. 2015. "New Games, New Rules, New Scoreboards: The Potential Consequences of Big Data," *Journal of Information Technology* (30:1), pp. 58-59.
Mattioli, M. 2014. "Disclosing Big Data," *Minnesota Law Review* (99), p. 535.
Metcalf, J. 2016. "Big Data Analytics and Revision of the Common Rule," *Communications of the ACM* (59:7), pp. 31-33.
Metcalf, J., and Crawford, K. 2016. "Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide," *Big Data & Society* (January-June 2016), pp. 1-14.
Narayanan, A., and Shmatikov, V. 2008. "Robust De-Anonymization of Large Sparse Datasets," *Security and Privacy, 2008*: IEEE, pp. 111-125.
Newell, S., and Marabelli, M. 2015. "Strategic Opportunities (and Challenges) of Algorithmic Decision-Making: A Call for Action on the Long-Term Societal Effects of 'Datification'," *The Journal of Strategic Information Systems* (24:1), pp. 3-14.
Nolan, R. L. 2012. "Ubiquitous It: The Case of the Boeing 787 and Implications for Strategic It Research," *The Journal of Strategic Information Systems* (21:2), pp. 91-102.
Otjacques, B., Hitzelberger, P., and Feltz, F. 2007. "Interoperability of E-Government Information Systems: Issues of Identification and Data Sharing," *Journal of Management Information Systems* (23:4), pp. 29-51.
PEW. 2013. "Conflict of Interest: Policies for Academic Medical Centers," Available at http://bit.ly/1U8M4Cp, last accessed by authors on 2/18/2017.
Scholtens, B., and Dam, L. 2007. "Cultural Values and International Differences in Business Ethics," *Journal of Business Ethics* (75:3), pp. 273-284.
Shaw, D. 2016. "Facebook's Flawed Emotion Experiment: Antisocial Research on Social Network Users," *Research Ethics* (12:1), pp. 29-34.
Venturini, T., and Latour, B. 2010. "The Social Fabric: Digital Traces and Quali-Quantitative Methods," *Proceedings of Future En Seine* (2009), pp. 87-101.
Yoo, Y. 2010. "Computing in Everyday Life: A Call for Research on Experiential Computing," *MIS Quarterly* (34:2), pp. 213-231.