Seeing the Forest for the Trees: Understanding and Implementing Regulations for the Collection and Analysis of Human Centered Data

Jana Diesner
The iSchool, University of Illinois at Urbana Champaign, Champaign, IL 61820, USA
jdiesner@illinois.edu

Chieh-Li Chin
The iSchool, University of Illinois at Urbana Champaign, Champaign, IL 61820, USA
cchin6@illinois.edu

Abstract

In human centered data science, researchers collect and analyze data about social entities. Some of the benefits of doing this are to gain substantive knowledge and develop new techniques and tools. The process of collecting data about individuals is regulated by multiple sets of explicit and implicit norms and rules, including personal ethics, Institutional Review Boards, privacy and security regulations, copyright, and terms of service. In this article, we outline why it can be challenging for scholars to keep track of all applicable rules, identify their meaning and compatibility, and practically implement them. We conclude that educational offerings and institutionalized processes need to be developed and implemented so that researchers can gain the awareness, knowledge, and skills essential for gathering and analyzing digital social trace data responsibly. We argue that scholars from the field of human centered data science need to be active participants in the public discourse and policy making on this topic since they can contribute domain expertise as well as methodological and technical insights.

Author Keywords
Ethics; digital social trace data; norms and regulations

ACM Classification Keywords
K.7.4. Professional Ethics

Introduction and Problem Statement

A large portion of human centered data science uses digital social trace data for analysis [13]. For this article, we define digital social trace data as information about people interacting with other social agents (e.g., via social networking platforms), pieces of information (e.g., on product-review sites and discussion forums), and socio-technical infrastructures (e.g., using phone apps) (for a more detailed definition of digital trace data see [10]). Since such data frequently can be gathered without researchers needing to interact with users (also called passive measurement [18]), and these data often are considered publicly available [19], an Institutional Review Boards (IRB) review might not apply. Once the applicability of an IRB has been ruled out, scholars might be insufficiently educated about additional rules and norms. These regulations depend on peoples' personal, cultural, and educational backgrounds as well as their institutional affiliations, among other factors. This article reviews reasons for this lack of expertise, the different rule sets that might be applicable when conducting responsible research, and reviews and proposes possible solutions.

In this paper, we focus on academic contexts; acknowledging that collaborations with other partners may involve additional challenges: Some companies are self-regulated by internal ethics review processes [9].

Many governmental and corporate organizations adhere to the "Fair Information Practice Principles" (FIPPs)[1]. Medical institutions must comply with the "Health Insurance Portability and Accountability Act" (HIPAA)[2] to protect patients' privacy. The Menlo report specifies privacy regulations for the field of computer and information security research regardless of institutional affiliation [5]. This continuously growing body of regulations related to privacy and ethics exists on top of what is discussed in this paper.

Why is this Problem not Solved?

We acknowledge that existing regulations might truly serve the purpose for which they were designed. For example, IRBs were developed "to protect the rights and welfare of humans participating as subjects in the research"[3].

Problem number one arises from the fact that some of these rules were defined in a pre-social media era and hence might not be easy to translate into the digital social data space. Efforts to catch up and align regulations with reality may lag behind technological advances and emerging practices.

Second, the lack of a clear understanding of rules beyond IRBs can be partially attributed to unclear, incomplete, or missing policies and regulations for working with digital social trace data [12].

---

[1] http://www.nist.gov/nstic/NSTIC-FIPPs.pdf

[2] http://www.hhs.gov/hipaa/index.html

[3] http://www.fda.gov/RegulatoryInformation/Guidances/ucm126420.htm

Third, research communities might have established practices that challenge or violate terms of service. This can occur either because those terms were changed and constricted over time and after initial publications of research studies, or because examples were set and adopted without considering all possible rule types [19]. For example, the terms of service of many product review sites and discussion forums do not allow or do not include permission to crawl or scrape data, but a recent study has shown that about 73% of the studied population (263 participants from academia, industry, the government and other sectors) think that it is permissible to "scrape data from online forums," and another 21% have a neutral opinion on this point [17]. Some peoples' understanding of this issue might be biased by the fact that some providers do allow data collection from their sites; though often via APIs instead of crawling or scraping. Once research that uses (now) debatable practices has been published, examples and quasi-standards are set, and consequently, students learning from these papers might make the wrong choices for the right reason [17].

Rules to Consider

What regulations exist? We have previously identified the following set (which may be incomplete, for details see [3]), and expand on a few points in this paper:

1. *Personal moral and ethics:* The Belmont report – one of the bases for IRBs – implements a set of ethics principles, namely respect, beneficence and justice. In addition, we argue that it is relevant to consider that each researcher – consciously or unconsciously - brings their own ethics to the table [18], which may or may not correlate with their training and field [17]. Also, ethics research has shown that people employ different moral principles [6; 16] depending on their gender [7], age [15], personal maturity [11] and culture [8; 16]. It might be naïve to assume that people put this part of their personality aside when planning their work, building their careers, or considering an externally defined rule set. Further research is needed to identify these processes and their implications.
2. *Institutional norms and expectations:* This includes IRBs, HIPAA, and data management plans[4], which are increasingly required by federal funding agencies, among others.
3. *Copyright and fair use:* Many webpages and apps leverage the fair use rule by only providing snippets of (appropriated) content (from other sources). Practically speaking, researchers might have lawful access to interaction data (e.g., which anonymized user replied to which comment from which other user), but no proper access to the full content of posts. This means that we might be able to collect social network data, but have to disregard substantial portions of natural language text data, even though it has been shown that network formation and language use mutually impact each other [2; 14]. Acting in a fully rule-compliant way might lead to losing this level of depth, comprehensiveness and rigor.
4. *Privacy laws and regulations*
5. *Security laws and regulations*
6. *Terms of service:* These are typically spelled out very explicitly, but might be difficult for people

[4] https://www.nsf.gov/eng/general/dmp.jsp

without a legal background to understand and translate into practical steps.

Seeing the Forest for the Trees: How to Arrive at Responsible Conclusions and Solutions?

Last year, as part of our work on assessing the impact of social justice documentaries on individuals, groups, and society, we were interested in analyzing customer reviews [4]. Many such reviews are publicly available. Also, using review data for social computing applications and building prediction models is common practice (we decided not to reference any specific studies). We asked three on-campus units for their advice on whether we can collect review data or not. Our IRB quickly determined that no IRB review or approval would be needed. Our library informed us that our work might or might not fall under the fair use rule. Our legal advisors discovered that the terms of service specified that crawling and scraping were not included or explicitly prohibited in the user agreements. In other words, we received three different opinions from the three stakeholders we consulted. As has probably happened to many other researchers before us, this uncertainty launched us into a process of trying to understand which rules do apply, and what these rules mean in practical terms. In the end, we devised an agreement with one review-data provider who allowed us to collect review data and use them for clearly defined purposes. There might be more efficient solutions to this problem; the fact that we do not know of any reflects the lack of clarity and standardized institutionalized processes or agencies that help with these questions and processes.

We suggest that both of these strategies – education and institutionalized processes - will need to be established so that scholars can gain the awareness, knowledge, and skills needed to find actionable answers and solutions for collecting and analyzing human centered data.

Summary

We acknowledge that it is no trivial task for researchers to 1) keep track of all the rules that are potentially applicable when working with human centered data, 2) identify their meaning and compatibility, and 3) practically implement them. What is the best way to prepare researchers for this task? We argue that universities are not yet sufficiently prepared to support scholars in this process, e.g. through educational offerings and governing bodies; both of which might be primary instruments to solve this issue. Some respective research [1] and initiatives are underway, and we expect more of this to come in the near future.

Finally, we as researchers and practitioners in human centered data science need to be active participants in the public discourse and policy making surrounding this issue. What we can provide are domain expertise on relevant research questions and experimental designs, and technical insights into computational solutions for data collection and analysis.

Acknowledgements

This work is supported by the FORD Foundation, grant 0155-0370, and a faculty fellowship from the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana Champaign.

References

1. Ryan Calo. 2013. Consumer Subject Review Boards: A Thought Experiment. *Stanford Law Review Online,* 97-102.
2. Steven R. Corman, Timothy Kuhn, Robert D. McPhee, and Kevin J. Dooley. 2002. Studying complex discursive systems: centering resonance analysis of communication. *Human Communication Research*, 28, 2: 157-206.
3. Jana Diesner and Julian Chin. 2015. Usable ethics: practical considerations for responsibly conducting research with social trace data. In *Proceedings of Beyond IRBs: Ethical Review Processes for Big Data Research*. Washington, DC.
4. Jana Diesner, Jinseok Kim, and Susie Pak. 2014. Computational impact assessment of social justice documentaries. *Metrics for Measuring Publishing Value: Alternative and Otherwise*, 17, 3.
5. David Dittrich and Erin Kenneally. 2011. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. *US Department of Homeland Security*.
6. Alan P. Fiske. 1991. *Structures of Social Life: The Four Elementary Forms of Human Relations: Communal Sharing, Authority Ranking, Equality Matching, Market Pricing*. Free Press.
7. Carol Gilligan. 1987. *Moral orientation and moral development*. In *Women and Moral Theory*, Eva Feder Kittay and Diana T. Meyers (Eds.). Rowman & Littlefield, 19-23.
8. Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101, 2: 366-385.
9. David Hoffman and Paula Bruening. 2015. Rethinking privacy: fair information practice principles reinterpreted. In *Proceedings of the 37th Annual International Data Protection and Privacy Commissioners' Conference*.
10. James Howison, Andrea Wiggins, and Kevin Crowston. 2011. Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, 12, 12: 767-797.
11. Lawrence Kohlberg. 1984. *The Psychology of Moral Development: The Nature and Validity of Moral Stages*. Harper & Row.
12. Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70, 6: 543-556.
13. David Lazer, Alex S. Pentland, Lada Adamic, Sinan Aral, Albert L. Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, and Myron Gutmann. 2009. Life in the network: the coming age of computational social science. *Science*, 323, 5915: 721-723.
14. James Milroy and Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics*, 21, 2: 339-384.
15. Jean Piaget. 1932. *The Moral Development of the Child*. Kegan Paul, London.
16. Richard A. Shweder, Nancy C Much, Manamohan Mahapatra, and Lawrence Park. 1997. The" Big Three" of morality (autonomy, community, divinity) and the" Big Three" explanations of suffering. In Morality and Health, Allan M. Brandt and Paul Rozin (Eds.). New York: Routledge, 119-169.
17. Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the Belmont principles: ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 9th ACM Conference on Computer-Supported*

Cooperative Work and Social Computing (CSCW). San Francisco, CA.

18. Bendert Zevenbergen, Brent Mittelstadt, Carissa Véliz, Christian Detweiler, Corinne Cath, Julian Savulescu, and Meredith Whittaker. 2015. *Philosophy Meets Internet Engineering: Ethics in Networked Systems Research. (GTC Workshop Outcomes Paper)*. Oxford Internet Institute, University of Oxford.
19. Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, 12, 4: 313-325.