Usable Ethics: Practical Considerations for Responsibly Conducting Research with Social Trace Data

Jana Diesner, Chieh-Li Chin (jdiesner@illinois.edu, cchin6@illinois.edu)

The iSchool at Illinois, University of Illinois at Urbana Champaign

Draft, Nov 2015

1. Introduction and Problem Statement

Over the last decade, research on the privacy of user information has shown that often a) ordinary users pay little attention to privacy policies and b) when considering policies, people have a hard time understanding their meaning and practical implications (Acquisti & Grossklags, 2005; Kelley, Hankes Drielsma, Sadeh, & Cranor, 2008; McDonald & Cranor, 2008). Usable computational solutions to this problem have been developed (Cranor, 2003; Kelley, Bresee, Cranor, & Reeder, 2009).

We observe a similar trend with respect to the ethics, norms and regulations for using public digital data at any scale; big and small. By this we mean that researchers may have little awareness of the different types of regulations beyond IRBs that might apply to their work, and difficulties to fully comprehend and implement applicable rules. This article focuses on practical issues with properly using social trace data for research and proposes solutions. For the purpose of this article, we define publicly available social trace data as information about people interacting with a) other social agents (e.g. social networks data from Facebook and Twitter), b) pieces of information (e.g. product review sites and discussion forums), and c) infrastructures (e.g. people checking in to places, geolocation services), (Howison, Wiggins, & Crowston, 2011), and natural language text data (e.g. the content of posts and tweets) that all can be collected without intervention or interacting with users (also called passive measurement (Zevenbergen et al., 2015)).

When planning a research project that involves intervention or interaction with living individuals, and/or identifiable, private information about individuals, at least university-based scholars in the US are required to have their intended work reviewed by their Institutional Review Board (IRB). The IRB will scrutinize any proposal for its compliance with basic ethical principles, such as respect for people, beneficence, justice, and minimization of risk (Research", 1979). However, an IRB review might not apply when gathering and analyzing public social trace data. An example would be passively observing and measuring internet traffic such as tweets, or comments on open discussion boards.

What happens once the need for an IRB review is ruled out? Are our ethical obligations taken care of, and we are ready to roll up our sleeves and start doing the work? Or are there other regulations, and when do they apply? In our day to day work, we observe that some graduate and post-graduate researchers might be insufficiently prepared to answer – or even ask - these additional questions. We argue that academe does not always provide or require the education that would be needed to equip the next generation of information professionals and data scientists with the awareness, knowledge and skills that would allow them to make informed decisions, solve problems and be responsible actors in data intense environments. This is a cross-disciplinary issue; it concerns students and scholars from engineering, computer science, the social sciences and humanities, public health, etc..

In this paper, we contrast different types of potentially applicable regulations beyond IRBs, relate them to different viewpoints - namely decision making primarily driven by ethical and privacy concerns, technical feasibility and other rule sets, and highlight practical implication of these viewpoints.

2. What Makes Conducting Responsible Research on Social Trace Data Difficult?

To motivate our argument, we provide some tangible examples for questions that we have encountered in our work in an Information Science environment, or that students have brought to us (the reader is cordially invited to see if they know the answers to these questions). The questions were generalized to provide an idea of the breadth of issues to consider. To keep this survey organized, we group those questions into four categories, which touch on different stages of the research cycle.

1. Data Access, Collection, and Storage:
    - Under what conditions are we permitted to collect what data from social media platforms and online data sites, e.g. social interaction data from Facebook or Twitter, discussion forums, customer review sites and online newspapers?
    - How can we collect the data? What’s the difference between APIs and scraping from a regulatory point of view?
    - How can we store the data?
    - Does fair use apply to code? To data? To word lists?
2. Data Use, Analysis, and Modification:
    - How can we anonymize social network data? The content of social media data?
    - What types of analysis can be run on social media data?
    - We found a suitable third-party ontology, lexicon or gazetteer for our project. How can I use it? Modify it? Share it? Use it for a commercial project?
3. Data Sharing and Publishing Research and Data:
    - Can we share the data with other stakeholders (academia, industry) collaborating in the same project?
    - We annotated some data, e.g. for a machine learning or digital annotation project. Under what conditions can we make what portions of the data available to others?
    - We wrote some cool new code and want others to be able to use. Who owns the copyright? How do we release the code? Do we need a license? Which one? Who issues the license, us or somebody at our organization? Can we request others to acknowledge our code? Does our organization need to be acknowledged? Can we be held responsible for issues that others experience due to using my code?
    - We want to publish our results in journal X, which has an open data policy. What does that imply for planning our study? Under what conditions can we release the data to a journal, the university library, a sponsor, or a project partner?
4. Project-level information management:
    - If an IRB does not apply to our project, is there an ethics or privacy review board, protocol or process?
    - What's the relationship between copyright, terms of service and usage/ privacy ethics? What trumps what?
    - We don't understand what the terms of service or use on a webpage mean in a practical sense. How do we develop this data science literacy?

- Last but not least: We got different answers to some of these questions from different stakeholders, e.g. the IRB, library, research services and legal counselors. How do we make an informed decision?

Having the awareness to ask these questions, and having the knowledge to answer them and the skills to develop solutions is not trivial. We believe that education can fix this problem, but related modules do not yet seem to be prevalent. Overall, when it comes to planning, executing and publishing research that involves digital data that were authored by humans or involve information about people, researchers might be unaware of, uneducated and/or confused about or overwhelmed by the set of rules that need to be evaluated for their applicability. To address this issue, we next provide and then discuss a brief summary of these rule sets (Table 1); acknowledging that there might be additional regulations that we are unaware of.

Table 1: Rules and pointers to units who might offer help with questions

Rule Set	Suggestions for who to ask for help
1. **Ethics**, which may depend on culture (Graham et al., 2011; Shweder, Much, Mahapatra, & Park, 1997)	Yourself and peers (developed over a lifetime, might be applied unconsciously)
2. **Norms and expectations**: Prior research has shown that the vast majority of adolescents and adults follow conventional morality, i.e. they comply with the norms at play in the groups they are embedded in. Only about 10-15% of adults develop post-conventional morality, i.e. they establish and are guided by their own ethical principles (Kohlberg, 1984).	Developed over a lifetime, might be applied unconsciously
3. Institutionally dependent and binding **rules**: E.g. IRBs, ethics acts, the Health Insurance Portability and Accountability Act (HIPAA[1], which is enforced by the Office for Civil Rights), or data management plans, which several federal funding agencies[2] requests as part of research proposals.	IRB, research services
4. **Privacy** regulations	IRB, research services
5. **Security** regulations	Technical services, helpdesk
6. **Copyright**, its modern variations, and fair use	Library
7. **Terms of service**/ use, **licenses**	Legal services

Groups 1 and 2 Table 1 in represent normative behavior, i.e. rules of conduct that individuals are expected to learn and adopt for interacting with others in their organizations and communities of practice (Wenger, 1999). Acquiring and executing this understanding can be difficult, e.g. when rules are tacit or dependent on culture (with the possibility of incompatibilities and conflict).

Groups 3 to 7 (which includes IRBs), for the most part, are laws or implement laws. Respective knowledge and skills can be taught. However, for many questions related to internet data, no laws exist yet, and trying to apply existing legislation might be a poor fit (Zevenbergen et al., 2015).

[1] http://www.hhs.gov/ocr/privacy/hipaa/understanding/special/research/index.html
[2] https://www.nsf.gov/eng/general/dmp.jsp

Table 1 furthermore shows that IRBs are only a small part of multiple bodies of concerns and regulations. However, for practical purposes, it can seem that IRBs are the rule set that gets most emphasized in graduate training programs. Also, the outlined rules might overlap and/or contradict each other. Furthermore, multiple sets might apply to a single research project. All of these points can complicate the proper conduct of research.

Different approaches to proper practical conduct of research exist. Relationship between the legality, ethics and technical feasibility for collecting and using various types of data and tools.

3. Practical Approaches to Working with Social Trace Data

a. Reasoning driven by technical feasibility

Driven by the technical feasibility of a project, researchers might sometimes take a fairly pragmatic approach to working with social trace data (Kosinski, Matz, Gosling, Popov, & Stillwell, 2015). One example would be to consider such data as a means to identify general patterns of social interaction and structure, as well as the underlying dynamics of socio-technical systems. Conceptualizing social data as a tool for obtaining generalizable knowledge about society, groups or communities as a whole (as opposed to its personally identifiable members) might explain why scholars are sometimes unaware or unconcerned about the ethics and societal implications of their work (Kosinski et al., 2015).

Some fundamental knowledge about the structure and functioning of the internet (Barabási & Albert, 1999; Newman, Barabasi, & Watts, 2006) and the evolution of links and groups is based on digital trace data (Backstrom, Huttenlocher, Kleinberg, & Lan, 2006; Tiropanis, Hall, Crowcroft, Contractor, & Tassiulas, 2015). Furthermore, knowledge gained this way is essential for designing and maintaining sustainable infrastructures and communities (Kraut et al., 2012). Kosinski and colleagues (2015) refer to analyzing public trace data is an instance of doing archival research, and suggest that participant consent is not needed if users consciously made their data pubic, collected data are anonymized, researchers do not interact with participants, and no identifiable user information is published.

When working with such data, as a starting point, it is important that scholars are aware of the fact that not everything that can be done should be done; i.e. technical feasibility doesn't translate into clearance for a project. Zevenberg and colleagues (2015) have associated reasoning from the point of view of technical practicability with utilitarian ethics (maximizing utility, "the ends justify the means"). However, once this awareness has been established, navigating the space of rules can be a daunting and confusing endeavor.

Some of the applicable terms of service for working with social media data are implemented in APIs that people running a platform provide. For example, using the Facebook API will not allow researchers to collect data from private groups. However, this approach only works if people do use APIs; scraping might circumvent these mechanisms. Also, more subtle questions might apply: If one uses a publicly available tool that leverage an API, e.g. NodeXL for collecting social media data (Hansen, Shneiderman, & Smith, 2010), but required people to log in with their credentials for data collection, how can a researcher guarantee to not tap into private data because one of their friends shared data with them that another researcher would not be able to see.

b. Reasoning driven by ethics

Today's diverse teams of researchers from different backgrounds, ethnicities and gender might also involve a variety of ethics. Decades of ethics and moral research have shown that people form their individual (gender-specific (Gilligan, 1987)) set of moral principles during their teenage years (Kohlberg, 1958; Piaget, 1932); i.e. once they become scholars, those principles might be deeply rooted in people.

Shweder (1997) suggests that people employ one or more of the following basic types of ethics: Autonomy driven people are primarily concerned with the protection of individual rights and justice. Community oriented minds prioritize the preservation of institutions and social order, which translates into a sense of duty, respect and loyalty. Finally, people concerned with divinity care to protect people's inner purity from degradation due to, for example, selfishness.

Fiske (1991) provides an alternative pluralistic approach of psychological bases of social life according to which people may be driven by rational self-interest (market pricing), a preference for sharing with others (communal sharing), deferring to authorities (authority ranking), and aiming for balanced benefits for oneself and others (equality matching).

Zevenberg and colleagues (2015) recently published a more in-depth discussion on the relationship between different types of ethics and approaches to working with online data. We acknowledge that diverse research teams might bring a variety of personal moral principles to the table that may translate into conflicting approaches to acquiring and using social trace data.

c. Reasoning driven by rule compliance

Ultimately, considering all bodies of rules that apply should lead to an appropriate solution for collecting and working social trace data. This process involves a variety of challenges:

- Comprehension: Most students outside of law schools might find it challenging to understand and interpret terms of use and service. For example, what does "not included in a license" mean in practical terms?
- Common practice: When planning a study, researchers – especially young students are more likely to learn from prior studies than from governing legal regulations. For example, scraping data from certain pages or platforms might have been permitted or not have been specified by terms of service, and people have used this technical solution to collect data and publish on their work. As terms of service or norms shift overtime, the published work might still set an example for others. Moreover, adopting best practices and following procedures within a community are cultural developments that might be hard to change.
- Lack of standards: Access and use of social trace data are not comprehensively regulated. To plug this hole, most commercial and other providers and hosts, e.g. Facebook, Amazon and Slashdot, have defined their own terms of service that complement given copyright and privacy regulations. It is up to the researcher to read and understand these regulations one by one, and to check for updates and changes of these rules.

Overall, many websites have provided terms of use and service to inform people about permitted data collection and use. To give an example, below we compare some rules for Facebook.com, Twitter.com, and Amazon.com (Table 2), and try to translate these rules into practical research implications (remainder of this section).

Table 2: Rules and practical behavior for research

Question	Facebook	Twitter	Amazon
Can I use the data for research?	**?** not mentioned	**YES** with restrictions	**?** the API's principle purpose is not research
Can I download/collect data?	**YES** need to a) obtain users' consents b) provide privacy policy	**YES** only through API	**YES and No** but not for benefit of third party
Can I use automated means (e.g. crawl, scrape, data mining) to collect data?	**NO** need prior permission of Facebook	**NO** need separate agreement with Twitter	**NO**
Can I analyze the data?	**?** not mentioned	**YES** with restrictions	**NO** need express prior written approval of Amazon
Can I modify the data?	**?** not explicitly mentioned, any use of data needs users' consents.	**NO** for security/privacy concern, contact Twitter in advance	**NO** except revising image or truncate text without altering meaning
Can I store/ carry the data on mobile/portable devices?	**YES** only on devices with associated authorized token	**?** not mentioned	**NO**
Can I share/redistribute/publish the data?	**NO** unless consents from users	**NO** unless prior written approval from Twitter	**NO** unless with express prior written approval of Amazon

a) Can I use the data for research?

According to the terms of service, some data from these sites are available for personal and non-commercial use, but it is not necessarily explicitly stated whether they can be used for research: currently, only Twitter mentions "research" in its rules and requests researchers to only provide lawful, nondiscriminatory and aggregated results. In addition, it is important to note that even though researchers might use the Amazon Product Advertising API to collect data such as customer reviews, the API is meant for "advertising and marketing the Amazon websites" rather than research. People need to submit an enrollment form to describe how the content obtained thought the API will be used for Amazon’s review and approval.

b) Can I download/collect the data?

These three websites allow people to download or collect data from their websites, mainly through APIs. Facebook requests people to obtain users' consent and provide Privacy Policy to explain how they are collecting and using the data. Twitter asks people to download data through their API. And Amazon requires people to not use data to benefit other third parties.

c) Can I use automated means to collect data?

Collecting data by scraping or crawling webpages is technically feasible unless blocked by the host. However, using automated means (e.g. crawling, scraping, data mining) to collected data is prohibited on these websites. People need to apply for permission from these websites. In addition, these services encourage people to use their APIs to collect data; in this way, the websites can control the scope and type of data that people can access and protect users' privacy.

d) Can I analyze the data?

For researchers, the main purpose with collected data is analyzing them (as opposed to building applications, for examples). Facebook does not explicit state whether people can analyze their data. Twitter allows people to perform analysis as long as it is lawful and nondiscriminatory and doesn’t identify a single person or small group of individuals. Amazon requires people to apply for prior written approval before analyzing their content.

e) Can I modify data?

Researchers sometimes need to modify or delete a portion of the data to protect people's privacy. These websites all have restrictions regarding modifying data. Modifying the format for display needs is usually allowed; however, modifying content needs prior consents from users or permission from the websites.

f) Can I carry the data on mobile or other portable devices?

For collaborative research and working with remote teams, it is common to upload data into cloud computing or storage services and accessing the data using different devices. These websites also have provide related regulations: Facebook requests people to only show data on devices associated with the authorized user access token. Amazon doesn't permit data to be used on applications that are intended for use with portable devices.

g) Can I share /redistribute /publish data?

All three websites require people not to share data unless they obtained prior permission from the websites and consents from the users. In addition, Twitter also requires people to share data using Twitter ID rather than sharing the original text from the content.

4. Discussion and Conclusion

Zooming out from the outlined regulatory and technical details to a macroscopic view of the impact of approaches to working with digital social trace data on science and knowledge discovery at large, we life in an era where a large portion of research about social systems is carried out by computer scientists (Kosinski et al., 2015). At the same time, questions that social scientists and anthropologists, for example, are eager to ask might not get answered because researchers from these fields might be more sensitive for social data issues or uncertain about technical solutions to them. Also, the unclear guidelines for the proper use of social data might slow down social scientists in their research due to lengthy IRB review processes, or – even worse – lead them to forego a project or modifying their research agenda altogether due to uncertainty and concerns about proper handling of privacy rights. This can have tremendous ramification for the evolution of our understanding of society. Moreover, computational approaches to working with interaction and text data often aim at modeling and formally describing social phenomena, such as identifying power law distributions of node degrees in a large variety of social networks (Barabási, 2003), while explaining the underlying psychological and social processes that lead to these effects might not get sufficiently attended to.

We conclude that education is needed to provide students and researchers with the knowledge and skills needed to select or design strategies and techniques for acquiring and using data and software, and sharing their results and other outcomes in a way that is compliant with the norms, laws, ethics and other types of regulations applicable in different practical domains. In order to make responsible and informed decisions, education modules would need to cover a three step process: First, awareness needs to be raised so that scholars understand: What bodies of rules might apply - beyond IRBs? Second, knowledge needs to be taught to help researchers develop and hone their literacy for rules and ethics. Third, people needs to acquire skills that let them translate their knowledge and best practices into actionable outcomes. Such training needs to cross the boundaries between computing, law, social science and philosophy, and be accessible to scholars from all of these fields, among others.

Acknowledgment

This work is supported, in part, by the FORD Foundation, grant 0155-0370.

References

Acquisti, A., & Grossklags, J. (2005). Privacy and rationality in individual decision making. *IEEE Security & Privacy*(1), 26-33.

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006, August 2006). *Group formation in large social networks: membership, growth, and evolution.* Proceedings of the KDD'06, Philadelphia, PA.

Barabási, A. (2003). *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. New York, NY, USA: Plume Books.

Barabási, A., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science, 286*(5439), 509-512.

Cranor, L. F. (2003). P3P: Making privacy policies more useful. *IEEE Security & Privacy*(6), 50-55.

Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing*: Free Press.

Gilligan, C. (1987). Moral orientation and moral development.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology, 101*(2), 366.

Hansen, D., Shneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL: Insights from a connected world*. Burlington, USA: Morgan Kaufmann.

Howison, J., Wiggins, A., & Crowston, K. (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.

Kelley, P. G., Bresee, J., Cranor, L. F., & Reeder, R. W. (2009). *A nutrition label for privacy.* Proceedings of the Proceedings of the 5th Symposium on Usable Privacy and Security.

Kelley, P. G., Hankes Drielsma, P., Sadeh, N., & Cranor, L. F. (2008). *User-controllable learning of security and privacy policies.* Proceedings of the Proceedings of the 1st ACM workshop on Workshop on AISec.

Kohlberg, L. (1958). *The development of modes of moral thinking and choice in the years 10 to 16*: University of Chicago.

Kohlberg, L. (1984). *The psychology of moral development: The nature and validity of moral stages* (Vol. 2): Harpercollins College Div.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist, 70*(6), 543-556.

Kraut, R. E., Resnick, P., Kiesler, S., Ren, Y., Chen, Y., Burke, M., . . . Konstan, J. (2012). *Building successful online communities: Evidence-based social design*: The MIT Press.

McDonald, A. M., & Cranor, L. F. (2008). Cost of reading privacy policies, the. *ISJLP, 4*, 543.

Newman, M., Barabasi, A., & Watts, D. (2006). *The structure and dynamics of networks*. Princeton, USA: Princeton University Press.

Piaget, J. (1932). The moral development of the child. *Kegan Paul, London*.

Research", T. N. C. f. t. P. o. H. S. o. B. a. B. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*: Office of the Secretary.

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The" Big Three" of Morality (Autonomy, Community, Divinity) and the" Big Three" Explanations of Suffering. In A. M. Brandt & P. Rozin (Eds.), *Morality and Health* (pp. 119-172).

Tiropanis, T., Hall, W., Crowcroft, J., Contractor, N., & Tassiulas, L. (2015). Network science, web science, and internet science. *Communications of the ACM, 58*(8), 76-82.

Wenger, E. (1999). *Communities of practice: Learning, meaning, and identity*. New York: Cambridge University Press.

Zevenbergen, B., Mittelstadt, B., Véliz, C., Detweiler, C., Cath, C., Savulescu, J., & Whittaker, M. (2015). Philosophy Meets Internet Engineering: Ethics in Networked Systems Research. (GTC Workshop Outcomes Paper): Oxford Internet Institute, University of Oxford

Amazon. (2012, November 21). Amazon.com Product Advertising API License Agreement. Retrieved from https://affiliate-program.amazon.com/gp/advertising/api/detail/agreement.html

Amazon. (2015, June 22) Conditions of Use. Retrieved from http://www.amazon.com/gp/help/customer/display.html/ref=footer_cou?ie=UTF8&nodeId=508088

Facebook. (2015, January 30). Statement of Rights and Responsibilities. Retrieved from https://www.facebook.com/legal/terms

Facebook. (n.d.) Facebook Platform Policy. Retrieved from https://developers.facebook.com/policy

Twitter. (2015, May 18). Developer Agreement. Retrieved from https://dev.twitter.com/overview/terms/agreement

Twitter. (2015, May 18). Developer Policy. Retrieved from https://dev.twitter.com/overview/terms/policy

Twitter. (2015, May 18). Twitter Terms of Service. Retrieved from https://twitter.com/tos?lang=en