THE COUNCIL FOR BIG DATA,
ETHICS, AND SOCIETY

October 2016

Improving Services—At What Cost? Examining the Ethics of Twitter Research at the Montana State University Library

Sara Mannheimer, Scott W. H. Young, and Doralyn Rossmann
Montana State University Library

Is research with Twitter data too good to be true?

Introduction

As social media use has become widespread, academic and corporate researchers have identified social networking services as sources of detailed information about people's viewpoints and behaviors. Social media users share thoughts, have conversations, and build communities in open, online spaces, and researchers analyze social media data for a variety of purposes—from tracking the spread of disease (Lampos & Cristianini, 2010) to conducting market research (Patino, Pitta, & Quinones, 2012; Hornikx & Hendriks, 2015) to forecasting elections (Tumasjan et al., 2010). Twitter in particular has emerged as a leading platform for social media research, partly because user data from non-private Twitter accounts is openly accessible via an application programming interface (API).

This case study describes research conducted by Montana State University (MSU) librarians to analyze the MSU Library's Twitter community, and the ethical questions that we encountered over the course of the research. The case study will walk through our Twitter research at the MSU Library, and then suggest discussion questions to frame an ethical conversation surrounding social media research. We offer a number of areas of

This case study was first written for the Council for Big Data, Ethics, and Society. Funding for this Council was provided by the National Science Foundation (#IIS-1413864). For more information on the Council, see: http://bdes.datasociety.net/.

ethical inquiry that we recommend be engaged with as a cohesive whole.

Social Media Community Analysis Research

Background

In 2012, as members of the MSU Library's social media team, we developed a strategy for building community through social media. The strategy focused first on studying and understanding our library's primary target community—undergraduate students. It then outlined a sustainability plan for creating and sharing content on social media related to the undergraduate student community.

Research Design

Our team designed a research study to investigate the effect of the social media strategy on the MSU Library's Twitter community (Young & Rossmann, 2015). The research involved collecting and analyzing Twitter data from followers of the MSU Library Twitter account. This data was collected via the Twitter API to build a dataset consisting of Tweets and user profile data from all 998 followers of the MSU Library's Twitter account. We analyzed user profiles and categorized them into nine user types, including business, librarian, student, faculty, community, and alumni. We analyzed Tweets from the MSU Library account and categorized them into sixteen content types, including event, blog post, book, database, and student life. And we analyzed interactions between the MSU Library Twitter account and other Twitter user accounts and categorized them into three interaction types: retweet, reply, and comment. Finally, we combined these analyses to create an "interaction rate" that could be measured by user type and post type.

MSU's Institutional Review Board (IRB) determined that our research did not pose a risk to human subjects and was therefore exempt from IRB oversight. The IRB classified our research under the following exempt category: "the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available, or if the information is recorded by the investigator in such a manner that the subjects cannot be identified, directly or through identifiers linked to the subjects."

Results, Application, and Dissemination

Analyzing the MSU Library's Twitter community allowed us to measure the effectiveness of our social media strategy. Our research revealed that certain user types engaged more often with certain content types. For example, users who were categorized as students demonstrated a high rate of interaction with content that was categorized as student life, but they demonstrated a low rate of interaction with database content. This analysis revealed that content and community were directly related. As a result of our research, the MSU Library social media team refined our social media content and engagement with an eye towards the MSU Library's target community of undergraduate students.

An article that we wrote reporting on the research—including screenshots of interactions with users—was published in an open access peer-reviewed journal (Young & Rossmann, 2015), and we posted a copy of the article in the MSU Library's Institutional Repository to encourage broad readership. We also presented these findings at national and international conferences, often referencing specific Tweets. We have not published the dataset of Tweets that we collected.

Background and Discussion Questions

Over the course of the Twitter research described above, questions arose suggesting that conducting research with social media data is not a simple act of mining public data. The ethical implications of the MSU Library's Twitter community analysis research can be examined from three related perspectives (van Wynsberghe, Been, & van Keulen, 2013; Alim, 2014; Mannheimer, Young, & Rossmann, 2016):

1. **Context**: In what context is the research being conducted?

2. **Expectation**: What are user expectations surrounding social networking services and the use of their social media data for research purposes?

3. **Value Analysis**: Does the benefit of the research outweigh the potential risks or privacy violations to social media users?

The background information and discussion questions below can help guide an ethical examination of our Twitter research at the MSU library.

1. Context

The Association of Internet Researchers states in its ethical guidelines that "rather than one-size-fits-all pronouncements, ethical decision-making is best approached through the application of practical judgment attentive to the specific context" (Markham & Buchanan, 2012, p. 4). The research described in this case study covers three distinct contexts: libraries, academic research, and Twitter.

Libraries

BACKGROUND

Libraries have a longstanding commitment to patron privacy. The American Library Association's Code of Ethics states, "We protect each library user's right to privacy and confidentiality with respect to information sought or received and resources consulted, borrowed, acquired or transmitted."[1] But as is evident from the wording of the Code of Ethics, the information that libraries are ethically bound to protect has traditionally meant information that the patron seeks or receives from the library. When patrons use social media, as Griffey points out, "some portion of the information being shared is being shared intentionally by the patron" (2010). Unless patrons directly engage with the library's Twitter account, patrons' Twitter data doesn't align with the traditional definition of patron data. Libraries are still in the process of developing policies to address the different types of patron data that results from 21st century technologies (Hess, LaPorte-Fiori, & Engwall, 2014). Until those policies are developed, social media data lies in an ethically murky space.

DISCUSSION QUESTIONS

1. How does Twitter data differ from traditional patron data?
2. Keeping in mind the library core value of patron privacy, is it ethical for librarians to analyze data from library Twitter followers in order to develop library services?

Academic Research

BACKGROUND

A key indicator of quality research is reproducible results. To enable reproducibility, the data used for analysis must be made available to other researchers. For the research described in this case study, we did not publish the Twitter data that we collected using the Twitter API. However, in order for our research to be reproducible, the data would have to be shared with other researchers. The ethics of releasing Twitter datasets to the public is unclear. Twitter data is governed by two legal structures: copyright law, and the Twitter Terms of Service. Current laws are ambiguous regarding what content is copyrightable on social media,[2] but Twitter's Terms of Service state that while users "retain [their] rights to any Content [they] submit, post or display," users also grant Twitter "a worldwide, non-exclusive, royalty-free license to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such

1. http://www.ala.org/advocacy/proethics/codeofethics/codeethics
2. https://www.lib.ncsu.edu/social-media-archives-toolkit/legal

Content."[3] As of August 2016, the Twitter Terms of Service state that anyone may use the Twitter API to "reproduce, modify, create derivative works, distribute, sell, transfer, publicly display, publicly perform, transmit, or otherwise use the Twitter Services or Content on the Twitter Services."[4] The Developer Policy furthermore instructs developers to "promptly respond to content changes reported through the Twitter API, such as deletions or the public/protected status of Tweets," and "Only surface Twitter activity as it surfaced on Twitter. For example, your Service should execute the unlike and delete actions by removing all relevant Content, not by publicly displaying to other users that the Tweet is no longer liked or has been deleted."[5] Indeed, research datasets collected via the Twitter API may include Tweets that have since been deleted by the user, or Tweets from users who have since either deleted their Twitter accounts or changed their account settings from publicly visible (Twitter's default setting) to protected, i.e. accessible only by a select group of followers.[6] The Developer Policy therefore suggests that researchers would have to update published datasets as Twitter content changes in order to comply with the policy—an unsustainable prospect.

In a 2014 blog post, developer Ed Summers of the Maryland Institute for Technology in the Humanities describes how he was able to publish Twitter data but avoid publishing Tweets that had been deleted or made private.[7] His strategy was to publish a collection of Tweets by providing only the Tweet IDs, along with instructions for "hydration"—a method for using Tweet IDs in order to access the associated Tweets in full on the Twitter site. However, if the "hydrated" dataset excludes Tweets that have been deleted or protected, the dataset will be different from its original form, thus hindering later reproducibility.

DISCUSSION QUESTION

Twitter users may delete Tweets at any time, and may switch their accounts from public to protected at any time. Tweets that have been collected for research purposes may not be publicly accessible via Twitter in the future. Therefore, a research dataset comprised of Tweets may change over time, which hinders reproducibility. Is it ethical to conduct Twitter research, knowing that this research is inherently unreproducible?

3. https://twitter.com/tos?lang=en#basicterms
4. https://twitter.com/tos?lang=en#restrictions
5. https://dev.twitter.com/overview/terms/policy#2.Update_Maintain_the_Integrity_of_Twitter
6. https://support.twitter.com/articles/14016
7. https://medium.com/on-archivy/on-forgetting-e01a2b95272

Twitter: Web Documents or Human Subjects?

BACKGROUND

MSU's Institutional Review Board (IRB) reviewed our study and determined that, while human subjects were involved in our project, our research posed essentially no risk to these subjects because it involved the collection and study of publicly available, existing data. Excluding publicly available data from human subjects oversight is a long-standing policy for IRBs, based on the assumption that publicly available data is inherently low risk to the subjects because any informational harm is already done by the publicness of the data. However, some data ethicists have noted that this once historically sound assumption is no longer accurate for big data research techniques (Metcalf & Crawford, 2016; Metcalf, 2016). Furthermore, most IRBs do not have policies that specifically address social media research. The argument for whether or not social media research should be overseen by IRBs hinges upon whether or not social media data is human subjects data (Solberg, 2010). Wilkinson and Thelwall (2011) argue that Twitter user data should not be considered human subjects data, but rather should be classified as web documents. This point of view continues to be widespread throughout the research community (Alim, 2014). However, as the amount of research being conducted using large Twitter datasets grows (Zimmer & Proferes, 2014), an increasing amount of research supports the idea that Twitter data should be classified as human subjects data. Human subjects data should be treated with additional care, and should be examined from the perspectives of informed consent and privacy (Beninger, et al., 2014; Gleibs, 2014; Rivers & Lewis, 2014).

DISCUSSION QUESTIONS

1. Was the IRB correct in its determination that this research uses publicly available, existing data? Why or why not?

2. Many IRBs currently lack guidelines regarding social media research. Without IRB to provide feedback and structure, what steps can be taken by social media research teams in order to proceed ethically?

Twitter: Informed Consent

BACKGROUND

The Belmont Report has long been the guiding standard for ethical research with human subjects. The report is structured around three principles: Respect for Persons, Beneficence, and Justice (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). Respect for Persons requires that "subjects enter into the research voluntarily and with adequate information." Beneficence can be simplified into two parts: "(1) do not harm and (2) maximize possible benefits and minimize possible harms." The principle of Justice concerns paying careful attention to the selection of research subjects.

The Menlo Report (Dittrich & Kenneally, 2012) takes the three basic principles from the Belmont Report and applies them to research regarding information and communication technology. Regarding informed consent, the Menlo Report states that decisions about informed consent "may be impacted by whether [researchers] have obtained valid authorization from their users – via explicit agreements or contractual terms of service – for participation in research activities" (p. 7-8). For the research described in this case study, it is not clear whether Twitter users agree to participate in research by agreeing to Twitter's Terms of Service. While Twitter's Terms of Service clearly inform users that their Tweets may be accessed and reused via the Twitter API, it does not explicitly state that Tweets collected via the Twitter API may be used for research purposes.

In a recent study, researchers conducted interviews with social media users to determine (1) user attitudes about social media data being used for research purposes, and (2) how well social media users understand this type of research (Beninger, et al., 2014). The resulting report provides a succinct summary of arguments for and against informed consent, from the user's perspective. The report places some responsibility on users to curate their own public content, and suggests that social networking services must clearly communicate about privacy and availability of user data. However, the report ultimately concludes that informed consent is necessary to nurture understanding between researchers and participants, and to ensure that the research complies with moral and legal requirements (p. 3).

DISCUSSION QUESTIONS

1. For research that uses Twitter data, do you think it is necessary to obtain informed consent from each Twitter user whose data is studied?

2. If researchers share their dataset publicly, is there a greater obligation to receive consent? If researchers share their dataset publicly, are they obligated to de-identify the data in some fashion (e.g., altering the Twitter users' handles)?

3. The Twitter API allows researchers to conduct social media research with thousands of users; such large-scale research makes it impractical to obtain informed consent from each individual user. Should social media research projects limit the number of users analyzed in order to make it possible to obtain informed consent from all users? Or can researchers do as Shilton and Sheridon (2016) suggest: focus simply on being "transparent with research subjects—in big or small studies—as a more engaged and meaningful form of informed consent" (p. 1917)? What are some ways that researchers using Twitter data can be more transparent with Twitter users?

4. When using Tweets from individual users in scholarly presentations and articles, should researchers contact these users to obtain permission to use their Tweets?

2. Expectation

BACKGROUND

Most users' Twitter accounts are visible to the public.[8] However, users may not expect their posts to be read and mined by those beyond their own community of followers. boyd (2014) has termed these liminal spaces between public and private "networked publics"—public spaces that are created through social media. boyd argues that many users assume that content posted on social media will be obscured by the huge scale of data available—just as a conversation in a public park is effectively private, although it could potentially be overheard by others nearby. The distinction between these networked publics and physical public spaces can be summarized in four key characteristics:

- "persistence: the durability of online expressions and content;
- visibility: the potential audience who can bear witness;
- spreadability: the ease with which content can be shared; and
- searchability: the ability to find content" (boyd, 2014).

Twitter's official position on Tweet privacy and Twitter data reuse are highlighted with block quote style "tips" within the document: "what you say on the Twitter Services may be viewed all around the world instantly. You are what you Tweet!" and "we encourage and permit broad re-use of Content on the Twitter Services. The Twitter API exists to enable this."[9]

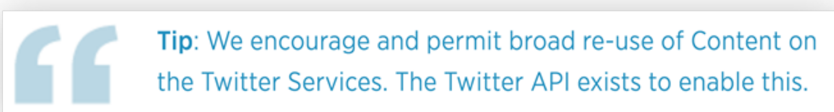

FIGURE 1 - Block-quote from Twitter's Terms of Service regarding Tweet Privacy

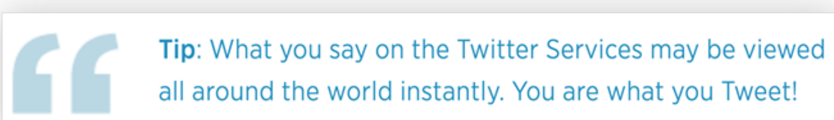

FIGURE 2 - Block-quote from Twitter's Terms of Service regarding Twitter data reuse

8. http://www.beevolve.com/twitter-statistics
9. https://twitter.com/tos?lang=en

These block quotes help users understand Twitter's Terms of Service at a glance. But research shows that most users do not read the online licensing agreements (Bakos, Mrrotta-Wurgler, & Trossen, 2014; Good, Grossklags, Mulligan, & Konstan, 2007; Böhme & Köpsell, 2010). In the case of Twitter, this means that users may not realize that their data is being made available to researchers through the Twitter API. Even if users read Twitter's Terms of Service, Twitter changes the document frequently,[10] and staying up to date can be difficult.

For the research described in this case study, the MSU Library website's Social Media page states that the library may reuse students' interactions with Library social media accounts "for research purposes and promotional materials so that we can understand and showcase our thriving online community."[11] Still, it is likely that many of the Library's Twitter followers have not read the Social Media page, and are therefore unaware that their data may be used for research purposes.

DISCUSSION QUESTIONS

1. How can researchers anticipate the expectations of Twitter users?
2. Do you think that Twitter users whose data are analyzed for research purposes know that their Tweets could potentially be used for research purposes?
3. How could the research design in this case study be adjusted to consider the expectations of the Twitter users whose data was collected?

3. Value Analysis

BACKGROUND

The Association of Internet Researchers ethical guidelines assert that all online research "must balance the rights of subjects (as authors, as research participants, as people) with the social benefits of research and researchers' rights to conduct research" (Markham & Buchanan, 2012, p. 4). The Twitter research described in this case study was designed to help the MSU Library build a sense of community using Twitter. Engaging with the community helps the library understand its users, and helps improve services to meet user needs. Creating a community-focused social media presence can also help encourage student well-being by nurturing a sense of belonging (Tomai, et al., 2010; Gray, Vitak,

10. https://twitter.com/tos/previous?lang=en
11. http://www.lib.montana.edu/about/social-media/#what

Easton, & Ellison, 2013; Yang & Brown, 2015). However, for the research described in this case study, we accessed and analyzed data from the members of the library's Twitter community without their knowledge or consent. We also highlighted non-anonymized Tweets from specific Twitter users in presentations and articles, without the knowledge or consent of those users.

DISCUSSION QUESTION

Do the benefits of the research described in this case study outweigh the risks? Why?

Conclusion

With its easily accessible API that allows researchers to download huge amounts of data, Twitter has the potential to be a large-scale source of insight into human opinions and behavior. It is particularly useful for public institutions such as libraries to have concrete data about how the public uses their services. But just because we can access this data, does that necessarily mean that we should? Thoroughly examining the ethics of a Twitter research project like the one conducted in this case study can help illuminate potential risks and benefits of research with Twitter data. Three related perspectives help frame this kind of ethical examination:

1. **Context**: In what context is the research being conducted?
2. **Expectation**: What are user expectations surrounding social networking services and the use of their social media data for research purposes?
3. **Value Analysis**: Does the benefit of the research outweigh the potential risks or privacy violations to social media users?

As the use of Twitter data for academic research becomes more widespread, it is increasingly important to continually discuss the ethical implications of this research. By applying structured ethical inquiry, we can place ethics at the center of future Twitter research.

References

Alim, S. (2014). An initial exploration of ethical research practices regarding automated data extraction from online social media user profiles. *First Monday, 19*(7). http://doi.org/10.5210/fm.v19i7.5382.

Beninger, K., Fry, A., Jago, N., Lepps, H., Nass, L., & Silvester, H. (2014). Research using social media: users' views. *NatCen Social Research*. Retrieved from http://www.natcen.ac.uk/media/282288/p0639-research-using-social-media-report-final-190214.pdf.

Bakos, Y. Marotta-Wurgler, F., & Trossen, D.R. (January 2014) Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts. *The Journal of Legal Studies 43*(1), 1-35. http://doi.org/10.1086/674424.

Böhme, R., & Köpsell, S. (2010, April). Trained to accept?: a field experiment on consent dialogs. In*Proceedings of the SIGCHI conference on human factors in computing systems*, 2403-2406. http://doi.org/10.1145/1753326.1753689.

boyd, d. (2014). *It's complicated: The social lives of networked teens*. New Haven: Yale University Press. Retrieved from http://www.danah.org/books/ItsComplicated.pdf

Dittrich, D., & Kenneally, E. (2012). The Menlo report: Ethical principles guiding information and communication technology research. *US Department of Homeland Security*. Retrieved from https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf.

Gleibs, I. H. (2014). Turning virtual public spaces into laboratories: Thoughts on conducting online field studies using social network sites. *Analyses of Social Issues and Public Policy, 14*(1), 352-370. http://doi.org/10.1111/asap.12036.

Good, N. S., Grossklags, J., Mulligan, D. K., & Konstan, J. A. (2007, April). Noticing notice: a large-scale experiment on the timing of software license agreements. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 607-616. http://doi.org/10.1145/1240624.1240720.

Gray, R., Vitak, J., Easton, E. W., & Ellison, N. B. (2013). Examining social adjustment to college in the age of social media: Factors influencing successful transitions and persistence. *Computers & Education, 67*, 193-207. http://doi.org/10.1016/j.compedu.2013.02.021.

Griffey, J. (2010). Social networking and the library. *Library Technology Reports, 46*(8), 34-37. Retrieved from https://www.journals.ala.org/ltr/article/view/4710/5605.

Hess, A. N., LaPorte-Fiori, R., & Engwall, K. (2015). *Preserving patron privacy in the 21st century academic library. The Journal of Academic Librarianship, 41*(1), 105-114. http://doi.org/10.1016/j.acalib.2014.10.010.

Hornikx, J., & Hendriks, B. (2015). Consumer Tweets about Brands: A Content Analysis of Sentiment Tweets about Goods and Services. *Journal of Creative Communications, 10*(2), 176-185. http://doi.org/10.1177/0973258615597406.

Lampos, V., & Cristianini, N. (2010, June). Tracking the flu pandemic by monitoring the social web. In *2010 2nd International Workshop on Cognitive Information Processing,* 411-416. http://doi.org/10.1109/CIP.2010.5604088.

Mannheimer, S., Young, S. W., & Rossmann, D. (2016). On the Ethics of Social Network Research in Libraries. *Journal of Information, Communication and Ethics in Society, 14*(2), 139-151. http://doi.org/10.1108/JICES-05-2015-0013

Markham, A. & Buchanan, E. (2012). *Ethical decision–making and Internet research: Recommendations from the AoIR Ethics Working Committee (version 2.0)*. Chicago, IL: Association of Internet Researchers. Retrieved from http://aoir.org/reports/ethics2.pdf.

Metcalf, J. & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society 3*(1), 1–14.

Metcalf, J. (2016). Big data analytics and revision of the common rule. *Communications of the ACM 59*(7), 31–33.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education and Welfare (1978). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research-the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research*. Washington, DC: US Government Printing Office. Retrieved from http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report.

Patino, A., Pitta, D. A., & Quinones, R. (2012). Social media's emerging importance in market research. *Journal of Consumer Marketing, 29*(3), 233-237. http://doi.org/10.1108/07363761211221800.

Rivers, C. M., & Lewis, B. L. (2014). Ethical research standards in a world of big data [version 2; referees: 3 approved with reservations]. *F1000Research 3*(38) http://doi.org/10.12688/f1000research.3-38.v2

Shilton, K., & Sayles, S. (2016, January). "We Aren't All Going to Be on the Same Page about Ethics": Ethical Practices and Challenges in Research on Digital and Social Media. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 1909-1918. http://doi.org/10.1109/HICSS.2016.242

Solberg, L. B. (2010). Data mining on Facebook: A free space for researchers or an IRB nightmare? *Journal of Law, Technology and Policy, 2010*(2), 311-342. http://ssrn.com/abstract=2182169.

Tomai, M., Rosa, V., Mebane, M. E., D'Acunti, A., Benedetti, M., & Francescato, D. (2010). Virtual communities in schools as tools to promote social capital with high schools students. *Computers & Education, 54*(1), 265-274. http://doi.org/10.1016/j.compedu.2009.08.009.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review, 29*(4), 402-418. http://doi.org/10.1177/0894439310386557.

van Wynsberghe, A., Been, H., & van Keulen, M. (2013). To use or not to use: guidelines for researchers using data from online social networking sites. http://doc.utwente.nl/87936/.

Wilkinson, D., & Thelwall, M. (2011). Researching personal information on the public web methods and ethics. *Social Science Computer Review, 29*(4), 387-401. http://doi.org/10.1177/0894439310378979.

Yang, C. C., & Brown, B. B. (2015). Factors involved in associations between Facebook use and college adjustment: Social competence, perceived usefulness, and use patterns. *Computers in Human Behavior, 46*, 245-253. http://doi.org/10.1016/j.chb.2015.01.015.

Young, S. W., & Rossmann, D. (2015). Building library community through social media. *Information Technology and Libraries, 34*(1), 20. http://doi.org/10.6017/ital.v34i1.5625.

Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management, 66*(3), 250-261. http://doi.org/10.1108/AJIM-09-2013-0083.