Data Is People: New Challenges and Tensions in Social Media Research

Jessica Pater, Parkview Research Center
Alicia Nobles, University of California, San Diego
Sarah Gilbert, University of Maryland
Michael Zimmer, Marquette University
Casey Fiesler, University of Colorado, Boulder

Duration:
1/2 day workshop

ABSTRACT:

Within social media research, a growing body of work blends social science and computational approaches. At the core of social media data are people creating and sharing content. This data is more pervasive and accessible than ever before. Many issues including fairness, transparency, and accountability as it relates to data collection, manipulation and assessment have sparked debate and thus created tensions and challenges within the field. This workshop aims to bring together a diverse set of participants to discuss how these tensions are being considered during the research design process, how researchers are handling these tensions and if these tensions are more prevalent in specific domains.

MOTIVATION

Research that blends social science and computational approaches is a diverse and growing discipline. The rapid evolution of technological advances has made data more pervasive and accessible than ever before. With this expansion, many tensions have arisen within the field.

Fairness, transparency, and accountability as it relates to the collection, manipulation, and assessment of the data are ongoing discussions, but often pose challenges for established mechanisms for ethical review. For example, in the United States, Institutional Review Boards (IRBs) regulate human subjects research--which the collection and analysis of publicly available data (e.g., tweets) is often not considered to be [7,9]. Therefore, computational research or even social science research that examines trace data (like much of the research published at ICWSM) does not have the clear-cut ethics guidelines (e.g., around consent) as some other types of research. However, trace data like tweets are still created by *people*, and represent both individuals and communities.

Often at the core of these tensions is protecting the people or communities in the data from identification or harm [6]. Recent scholarship in big data and research ethics has highlighted the diversity of domains where these tensions have increasing visibility. For example, the tension between individual privacy and usage of public data highlight complexities like the often ambiguous boundary between public data [8], reverse identification of data [1], identification of people [2], and data labeling (e.g., identifying the sexuality or mental health of a poster) [5,8]. Underlying this tension is the potential misalignment between the perception of the people who create the data and those that use it for secondary analyses [2]. Additionally, the tension between privacy protections and open science is growing as the nature of the data collected for research is increasingly more personal while fields are simultaneously promoting and expecting datasets to be published in open repositories [3].

Based on these considerations, when is it appropriate to contemplate these or other tensions or challenges in social media research (e.g., apriori to the research or post hoc case studies)? Are there topical domains where these considerations are more prevalent (e.g., for stigmatized issues like mental health)? How are researchers handling these tensions? This workshop aims to convene a dynamic and diverse subset of the ICWSM community to discuss their data collection, analysis, assessment, and reporting practices related to these tensions.

WORKSHOP STRUCTURE

The proposed ½ day workshop will follow the following structure:

- Welcome and overview by workshop organizers
- Brief participant introductions + Position paper lightning talk
- Short break
- Group Exercise 1
- Discussion
- Wrap-up
- Post-workshop networking.

The actual duration of the proposed activities will be adjusted based on the total number of workshop participants and the number of formal position papers that are submitted to the workshop.

Position Paper Lightning Talks

Participants that submit formal position papers will be given the opportunity to present to the full workshop.All presentations will be limited to 3-5 minutes each. After the completion of all lightning talks, workshop participants will have time to ask further questions of all presenters.

Workshop Exercise

Workshop participants will be assembled into groups and presented with a set of hypothetical research tensions involving social media research. They will be given a brief in which they will be asked to do the following and then discuss at the conclusion of the workshop:

- Outline the tension(s) presented in the brief.
- Provide a list of considerations related to the tension(s).
- Propose how researchers might address these tension(s).
- What open questions remain related to the tension(s) presented?

Below are examples of research tensions that will be presented to the workshop participants. All tensions will be emailed to participants prior to the workshop so they have time to consider the nuances and complexities of the proposed tensions.

1. Researchers plan to scrape public comments from online newspaper pages to predict election outcomes. They will aggregate their analysis to determine public sentiment. The researchers don't plan to inform commenters, and they plan to collect potentially-identifiable user names. Scraping comments violates the newspaper's terms of service.

2. Researchers plan to scrape profile photos, which are visible to any member of the service, from a dating site to build models that predict sexual preference or behavior. Researchers will not inform the dating site users, but they will not collect any identifying information and their photograph dataset will not be released publicly. Creating a fake profile, necessary to access the photos, violates the dating site's terms of service.

3. Researchers plan to scrape public posts and interactions from Facebook to study group-level dynamics. They plan to collect informed consent from the original poster, but not those they

interacted with, and they may collect identifying information. Scraping posts with permission of the original poster does not violate Facebook's terms of service.

4. Researchers plan to scrape data from an open health forum and combine it with scraped tweets to predict mental health conditions. The researchers will not inform forum users, and they may collect potentially identifying information. Scraping data violates neither the health forum nor Twitter's terms of service.

Post-workshop Networking

The organizers will setup a networking event for individuals participating in the workshop that is free to participants. We will work with the workshop chairs to ensure that our networking event does not conflict with other scheduled opportunities for people participating in the workshop. Refreshments will be provided.

SUBMISSIONS

We will accept submissions in the form of either formal position papers or lightweight statements of interest. Submissions should be no more than 4 pages (no minimum) and can be formatted in any style (though please submit a PDF). Submissions should be related to challenges or tensions in research using social media data, including but not limited to:

1. studies or works-in-progress;
2. description of a particular approach to ethics, supported by your or others' work;
3. cost-benefit analyses of particular research;
4. case studies of data tensions faced in your own work;
5. any statement of interest on the subject matter and why you would like to participate in the workshop

AAAI has discontinued publishing workshop proceedings, so accepted submissions will be shared only on the workshop website (or only with workshop participants, at the authors' request).

Our intention is to make this workshop as inclusive of different ideas and experiences as possible! If you are interested in all at having conversations about these issues but do not have the bandwidth or material for a full position paper, please consider sending in a statement of interest.

Information about the workshop, accepted submissions, and information about the organizers can be found at www.sites.google.com/view/dataispeople.

All submissions should be emailed to dataispeopleworkshop@gmail.com.

RELATED WORKSHOPS

Several of the workshop organizers have been involved in developing related workshops focused on discussing aspects of the tensions discussed which were couched in research ethics. Fiesler was the lead organizer of the "Ethics for Studying Sociotechnical Systems in a Big Data World" workshop at CSCW 2015 and a co-organizer for the "In Whose Best Interest?: Exploring the Real, Potential, and Imagined Ethical Concerns in Privacy-Focused Agenda" workshop at CSCW 2017. Fiesler and Pater were co-organizers of "Ethics and Obligations for Studying Digital Communities" workshop at GROUP 2016 and the "Challenges and Futures for Ethical Social Media Research" workshop at ICWSM 2016. Fiesler and Zimmer co-organized the "Exploring Ethical Trade-Offs in Social Media Research" workshop at ICWSM 2018.

To encourage participation, instead of exclusively focusing on the ethical aspects of research using social media data, this workshop takes a generalized approach in highlighting and workshopping ideas on how to address tensions related to social media research, which could include ethical issues.

ORGANIZERS:

Jessica Pater is a Research Scientist at Parkview Research Center. Her research examines how mobile and social technologies influence mental health and how technology use of patients connects to clinical practice. She was recently selected to serve on the ACM SIGCHI research ethics committee. She has helped organize various workshops focusing on ethics and research with vulnerable populations for CSCW, GROUP, and ICWSM.

Alicia Nobles is an assistant professor in the Department of Medicine at the University of California San Diego. She holds a PhD in Systems and Information Engineering. Her research examines how people seek information and support online for stigmatized health issues and uses these insights to build and deploy responsive public health resources. She lectures on social media research ethics in public health.

Sarah Gilbert is a postdoctoral research scholar at the University of Maryland College Park on the PERVADE: Pervasive Data Ethics project. She holds a PhD in Library and Information Studies. She studies policies and practices that make online communities healthier, including topics like what influences participation, how people learn in online communities, and how volunteer moderators' labor impacts community governance.

Michael Zimmer is an Associate Professor in the Department of Computer Science at Marquette University. He is a privacy and internet ethics scholar, whose work focuses on digital privacy, the ethical dimensions of social media & internet technologies, and internet research ethics. Dr. Zimmer is a co-chair of the Association of Internet Researchers (AoIR) Ethics Working Group, and a principal investigator in the PERVADE: Pervasive Data Ethics project.

Casey Fiesler is an assistant professor in the Department of Information Science at the University of Colorado Boulder. She holds both a law degree and a PhD in human-centered computing, and her research focuses largely on forms of governance online, including social norms, law, and ethics. She has organized a series of research ethics workshops at conferences including CSCW, GROUP, and ICWSM. She is a member of the ACM SIGCHI research ethics committee.

REFERENCES

[1] Ayers, J.W., Caputi, T.L., Nebeker, C. *et al.* Don't quote me: reverse identification of research participants in social media studies. *npj Digital Med* 1, 30 (2018). https://doi.org/10.1038/s41746-018-0036-2

[2] Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. Soc. Media Soc. 4, 1 (January 2018), 2056305118763366.

[2] Dennis, S., Garrett, P., Yim, H. *et al.* Privacy versus open science. *Behav Res* 51, 1839–1848 (2019). https://doi.org/10.3758/s13428-019-01259-5

[3] Hauge, MV, Stevenson, MD, Rossmo, DK (2016) Tagging Banksy: Using geographic profiling to investigate a modern art mystery. Journal of Spatial Science 61(6): 185–190.

[5] Sam Levin. 2017. New AI can work out whether you're gay or straight from a photograph. The Guardian.

[6] Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in big data research? The emerging ethics divide. Big Data Soc. 3, 1 (2016), 1–14.

[7] Jacob Metcalf. 2017. “The study has been approved by the IRB”: Gayface AI, research hype and the pervasive data ethics gap.

[8] Anja Thieme, Danielle Belgrave, Aane Sano, Gavin Doherty. 2019. Reflections on mental health assessment and ethics for machine learning applications. ACM Interactions.

[9] Jessica Vitak, Nicholas Proferes, Katie Shilton, and Zahra Ashktorab. 2017. Ethics Regulation in Social Computing Research: Examining the Role of Institutional Review Boards. J. Empir. Res. Hum. Res. Ethics (August 2017), 1556264617725200.

[10] Michael Zimmer. OkCupid Study Reveals the Perils of Big-Data Science. WIRED. Retrieved April 23, 2018