Research ethics in the Facebook era: privacy, anonymity, and oversight

Nathan Bos
Johns Hopkins Applied Physics Laboratory
11100 Johns Hopkins Rd
Laurel, MD 20723 USA
nathan.bos@jhuapl.edu

Karrie Karahalios
University of Illinois at Urbana-Champaign
Siebel Center 201 N. Goodwin Ave.
Room 3110
Urbana, IL 61801
kkarahal@cs.uiuc.edu

Marcela Musgrove-Chávez
University of Illinois at Urbana-Champaign
501 E. Daniel St
Champaign, IL 61820
mmusgrove@gmail.com

Erika Shehan Poole
School of Interactive Computing
Georgia Institute of Technology
85 5th St NW
Atlanta, GA 30332-0760
erika@gatech.edu

John Charles Thomas
IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY
jcthomas@us.ibm.com

Sarita Yardi
School of Interactive Computing
Georgia Institute of Technology
85 5th St NW
Atlanta, GA 30332-0760
syardi3@gatech.edu

CHI 2009, April 4 – 9, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-247-4/09/04.

Abstract
Ethical standards for human subjects research have not kept up with new research paradigms. Several research areas are particularly problematic for the CHI community. Online social research is testing the boundaries of public observation, third-party disclosure, and anonymization methods. Furthermore, there are differences in norms about what is and is not ethical among various research disciplines studying the Web. This SIG brings together members of the CHI community who are interested in research ethics for studying the Web. We invite seasoned veterans from industry and academia, educators, and newcomers to the field to share their experiences and advice, ask questions, and to form an interest group that can help shape university and corporate best practices for online research.

Keywords
Research ethics, regulation, methods, web research, social and legal issues

ACM Classification Keywords
K.4.1 Public Policy Issues, K.7.4 The Computing Profession

Human subjects research has been closely regulated in the United States since 1974, after a number of egregious violations of research ethics came to light, such as the Tuskeegee syphilis study, in which study participants suffered the effects of syphilis untreated for decades after effective treatments were known. In response to this and other events, Health and Human Services put forth Title 45 CFR 46 of Federal Regulations in 1975 [7]; the Belmont Report [8] outlining principles for ethical human subjects research was issued in 1979. Together, they provide the basis for ethical oversight of federally-funded human subjects research in the United States. Other countries, notably the United Kingdom and Australia, also have similar regulatory oversight over research.

While the original focus of these regulatory mechanisms involved medical studies, these same regulations guide *other* types of studies not initially envisioned by the creators of 45 CFR 46 or the Belmont Report, including online studies conducted within the CHI community. As a result, there is much that these regulations and guidelines do not answer in this context. This SIG brings together researchers interested in discussing the challenges of how to conduct ethical online research.

Online studies in particular create new research opportunities, but also introduce unforeseen ethical challenges related to privacy, anonymity, and differing norms. CHI researchers are increasingly finding themselves in unfamiliar territory. In some cases, existing standards are poorly aligned to emerging research paradigms. Online research spans geographic boundaries and disciplinary domains, yet has broadly disparate approaches and methods. As a community, we are at crossroads in determining what types of research activities are ethical with respect to online data, and how we can integrate these approaches into HCI education. As HCI researchers, what should our position be on research ethics and the role of external regulatory bodies such as Institutional Review Boards (IRB) in the type of work we do? What are the ethical issues in measuring and collecting data on the Web? What regulatory standards exist for governing research online? What pedagogical approaches should SIGCHI adopt for training students to do ethically sound research online?

Topic 1: Web data collection and analysis

Ethics in online research is not a new topic (*e.g.* [1]). However, the rapidly expanding scale and broadening scope of types of online research in cross-disciplinary communities demands attention within the CHI community. While researchers and ethics review boards often rely on dichotomies like "public" versus "private," "published" vs. "unpublished," and "anonymous" vs. "identified" [3], these categories are imprecise. We highlight two particular topics of concern: consent issues that are unique to online research, and limitations of data masking techniques for Web datasets.

First, consent is a particularly tricky issue in online research. On the Internet, the boundaries between public and private are not always clearly defined [1]. In real-world public settings, activities such as observation of a public village square is, by most accounts, considered an ethically sound research approach. But in an online setting, these boundaries are less clear. In what contexts should researchers seek consent from participants? When, and by what means, should people who use online services be

notified that online research is occurring? When, if at all, should anonymization or masking of data be required? Does data that is obtainable online automatically count as "public"?

We have little precedence and few clear guidelines for when these types of activities are appropriate. Terms of service, for example, vary across sites, and are governed by both legal policy (*e.g.* COPPA) as well as corporate policy. Further, online social networking research cannot be reasonably conducted, in many cases, if consent must be obtained from every potential data contributor. How should potential study participants be notified that their data may be used for research? What are the prerequisites for a weaker form of disclosure or 'opt-out' technique?

More recently, researchers have been studying temporal and ephemeral online phenomena, such as site failures (*e.g.* Google's unintended malware notice) and disaster relief (*e.g.* Twitter use during fires). These rapid and often large-scale episodes can be technically difficult to capture real-time, and often prohibitively difficult to capture if IRB (or equivalent) approval must first be gained before gathering data. Other disciplines have alternate ways to conduct "unanticipated" research. What steps can we take as a research community to develop protocol and best practice for conducting this unanticipated research?

Finally, current practices for anonymizing large datasets may not be adequate for many types of online research. Even if researchers take all reasonable steps to anonymize large datasets from the Web, it is possible that one's identity can be revealed. For example, researchers at UT Austin developed a de-anonymization technique for Netflix to show how an adversary who knows only a little bit about an individual subscriber can identify this subscriber's record in the dataset [11]. In this case, the research intentions were benevolent, but the method exploited undefined terrain in terms of human subjects. Other techniques that can identify individuals based on data such as search result histories or writing samples, making the process of anonymization more difficult than simply removing names from data.

Topic 2: Disparate norms across research communities

Many universities, government research labs, and private organizations maintain committees (*e.g.* Institutional Review Boards) to oversee human subjects research conducted by their organizations. Yet, rules and regulations vary across regions and universities. Given the diversity of SIGCHI's membership, which includes HCI researchers worldwide from both industry and academia, what should its position be toward conducting ethical online research?

Moreover, as online research expands into a broad cross-section of disciplines, there is little agreement on how to conduct ethically sound research. In studies of social networks, for example, researchers in computer networking and data mining study sites like Twitter and Facebook (*e.g.* [5,6]), yet the methods and approaches are distinctly domain-specific and look quite different to similar studies within the CHI community (*e.g.* [4]). Similarly, the role of site policies and terms of service are weighted differently among different academic communities and cultures. How can SIGCHI engage with other communities within ACM (*e.g*. SIGCOMM) who are also interested in online research?

Topic 3: Pedagogical approaches and issues in teaching ethics

Online research is increasingly been taught across disciplines and departments in higher education courses. Some courses are focused on network theory and analysis, while others are focused on the design and community aspects of the Web. Some include a mandatory ethics lecture and a class-wide IRB protocol under which students can then conduct individual projects [1,2]. However, as these types of courses pervade university classes across the country, there is a need for consistent pedagogical approaches to teaching students ethics in the context of online communities research. How should HCI educators train students (including both future HCI researchers as well as those who may not become researchers) to deal with ethical issues in online research? What should students be taught about when and how to capture, analyze, and report Web data?

Expected Outcome

This SIG will create a network of researchers who share challenges in conducting research online. Our goals are to generate critical discussions of the role of human subjects in online research in the CHI community; issues and stories of organization-specific challenges will be saved for a post-SIG discussion board. We expect to produce: a list of emerging topics in online research; a body of case studies for future benefit of the CHI community; a provocative examination of concerns related to policy, terms of service, privacy, anonymization, and research community norms in online studies; and a body of approaches to tackling emergent issues in online research.

References

[1] Bruckman, A., "Studying the amateur artist: A perspective on disguising data collected inhuman subjects research on the Internet," Ethics and Inf. Tech. 4, 3 (2002), 217-231.

[2] Bruckman, A., "Teaching Students to Study Online Communities Ethically," Journal of Information Ethics 15, 2 (2006), 82-98.

[3] Hudson, J.M. and Bruckman, A. Using empirical data to reason about internet research ethics. In *Proc. CSCW '05*, ACM, (2005).

[4] Joinson, A.N. Looking at, looking up or keeping up with people?: motives and use of facebook. In *Proc. CHI '08*, ACM, (2008).

[5] Kleinberg, J.M., "Challenges in mining social network data: processes, privacy, and paradoxes," In *Proc. 13th ACM SIGKDD*, ACM (2007).

[6] Nazir, A., Raza, S. and Chuah, C.-N. Unveiling facebook: a measurement study of social network based applications. In Proc. SIGCOMM, ACM, (2008).

[7] U.S. Department of Health and Human services. Title 45 Public Welfare, Part 46 Protection of Human Subjects. (Retrieved October 3, 2007) http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm

[8] U.S. Health & Human Services. The Belmont Report. (Retrieved October 3, 2007) http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm