Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community

Jessica Vitak
College of Information Studies,
University of Maryland
College Park, MD USA
jvitak@umd.edu

Katie Shilton
College of Information Studies,
University of Maryland
College Park, MD USA
kshilton@umd.edu

Zahra Ashktorab
College of Information Studies,
University of Maryland
College Park, MD USA
parnia@umd.edu

ABSTRACT

Pervasive information streams that document people and their routines have been a boon to social computing research. But the ethics of collecting and analyzing available—but potentially sensitive—online data present challenges to researchers. In response to increasing public and scholarly debate over the ethics of online data research, this paper analyzes the current state of practice among researchers using online data. Qualitative and quantitative findings from a survey of 263 online data researchers document beliefs and practices around which social computing researchers are converging, as well as areas of ongoing disagreement. The survey also reveals that these disagreements are not correlated with disciplinary, methodological, or workplace affiliations. The paper concludes by reflecting on changing ethical practices in the digital age, and discusses a set of emergent best practices for ethical online research in social computing.

Author Keywords

Ethics; big data; privacy; researchers; academia; online data; social computing

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION

In a well-known 2008 study [21], researchers shared "a new public dataset based on manipulations and embellishments of a popular social network site, Facebook.com." This comprehensive, longitudinal dataset of student networks and activities offered new possibilities for studying the dynamics of college relationships. Within days of the data's release, however, outside researchers re-identified the university and some individuals. As noted by Zimmer [38], the authors failed to understand *why* this re-identification was problematic, revealing a lack of knowledge of both technical and ethical issues in their research.

Seven years later, researchers still struggle to balance research ethics considerations with the use of online datasets. A study published in 2014 by Facebook and academic researchers [19] received significant public criticism for perceived ethics violations, including a lack of informed consent and potential risks to research subjects [11]. This recent study was a catalyst for public discussion about research ethics, amplifying a conversation active among Internet researchers for the past decade [7,24]. Internet researchers increasingly recognize that collecting digital trace data creates challenges for ethical codes developed last century, most notably the Belmont Report [29] and resulting Common Rule legislation. Numerous researchers [13,20,33,35], working groups [24], and workshops [9,10,39] have begun exploring this tension.

This paper adds to the conversation by describing the current state of online research ethics beliefs and practices in social computing. The paper documents social computing researchers' current online data collection and analysis practices; the challenges they experience; and their beliefs about what constitutes ethical online social computing research. Qualitative and quantitative analyses provide data that identifies areas of consensus and disagreement in the field. In the following sections, we present background on the specific challenges faced by online researchers, then present analyses from a survey of 263 academics, industry professionals, and other researchers who use online data. The findings represent one of the first studies of *researchers'* ethical considerations when collecting and analyzing data collected from online sources.

This paper also sheds light on what has been cited as a major challenge for online research ethics: *creating standards for fields with different backgrounds, methods, and approaches* [24]. Data from researchers in the social, information, and computer sciences illustrate that there are surprisingly few significant differences in ethical beliefs or practices between fields. Rather, disagreement on a few fundamental issues occurs across all fields. This study highlights those issues as challenges on which we need to build future consensus. It also provides an initial set of heuristics to further discussions of ethical practices in online data research.

ETHICAL CONCERNS IN ONLINE RESEARCH

Increased public participation in online networked spaces, along with the relative ease of collecting online traces, activities, and disclosures has accelerated research involving collection and analysis of data about users [17].

Searches of ACM and IEEE digital libraries find thousands of authors engaged in research using social media, online forums, trace ethnography, text mining, or activity traces.

Many forms of publicly available, pervasive, and "big" digital data used in such studies raise challenges to traditional definitions of core research ethics principles. Research in the U.S.[1] has been governed for more than 30 years by principles established in the Belmont Report, written in response to revelations regarding unjust studies in medicine and psychology [25]. The report emphasized three guiding principles: respect for research participants, beneficence, and justice in participant selection. Subsequent legislation based on the report, known as the Common Rule, codified these principles in the establishment of institutional review boards (IRBs), which oversee research ethics on U.S. campuses receiving federal funding [25].

Respect, beneficence, and justice remain relevant guiding principles for online data research. But collecting and analyzing online data often challenges the traditional interpretations of these principles by the Common Rule and IRBs. For example, *informed consent*—one way IRBs interpret *respect for persons*—is difficult to obtain from individuals contributing content to massive, publicly available information streams. Beneficence has frequently been interpreted as weighing relative risks against the benefits of research. Risk to participants is often minimized in human subjects research through confidentiality or anonymity, but these protections are both challenged in online data collection [30]. In addition, the risks to participants of exposure due to identification of online data are not well understood. Finally, fairness in research subject selection, long a tenet of *justice*, is challenged by the uneven demographic use of social media platforms [1,5]. We explore each of these challenges in more depth below.

Challenges to Respect for Persons

Informed consent challenges are one of the most frequently discussed topics in online research ethics. Online research subjects are sometimes unaware of monitoring, and often unable to choose the kind of data collected [6]. Online data subjects also have uneven opportunities to protect their data. While individuals increasingly use privacy settings provided by social networks [22], researchers allied with host platforms may still have access to the data. Transparency is another challenge for social computing researchers. While social media's affordances simplify the process of collecting data, researchers must decide whether and how to inform subjects of their presence, methods, and analysis. For example, a study that explored users' willingness to respond to strangers' questions on social media used Twitter handles like @TSAtracker to ask users about the wait time at airports. This handle masked the presence of researchers, introducing a degree of deception into the research [27].

Challenges to Beneficence

Balancing risks and benefits to participants and society is another challenge for online data research. One traditional tool for protecting individual research participants from harm has been protecting their identity. Ensuring the anonymity of participants in online data collection, however, is not an easy task. Studies have proven that anonymized data can be de-anonymized when paired with other datasets [3,23,26,30,38]. Minimizing risks to subjects is further complicated by the fact that the risks of exposing information about online research subjects are not well understood. For example, social media data can reveal information about the mental health of an individual [8]. The potential for misuse of this information is high, but such data are potentially useful for detection and treatment.

Challenges to Justice

Online data also present challenges to fairness or justice. Online participation does not mirror U.S. or global demographics [1,12]. Demographic variables such as race, gender, age, socioeconomic status, and technical experience impact individuals' self-selection into social network sites and online forums [12]. These differences can lead to biased samples when collecting data from social media sites. This point is highlighted in critiques of a 2015 Facebook study on exposure to ideological content [2]. Critics cite a sampling framework that limited participation to active users who self-disclosed their ideological affiliation (see [34,37]). This sampling frame limited the participant pool to just 4% of American users, significantly limiting generalizability of findings beyond the sample. A second challenge is that researchers do not have uniform access to online data. Researchers allied with commercial platforms frequently have access to more data than unallied researchers, leading to differential access among industry and academic researchers [6].

In summary, researchers face a number of barriers to conducting online research that aligns with the principles of the Belmont Report as implemented in the Common Rule. Below, we unpack how social computing researchers navigate these challenges.

METHOD

To better understand existing ethical norms in the interdisciplinary community conducting research with online data about people, we used the online survey platform SurveyGizmo to collected survey data from 263 researchers who self-identified as working with online data.

Survey Development and Deployment

We developed a survey to elicit the *ethical challenges* online researchers face, their *current practices* to respond to

[1] Research ethics protocols vary worldwide. European Union member states report to Research Ethics Committees (RECs), while Australian researchers follow a Code for the Responsible Conduct of Research. Both enforce policies similar to those in the U.S. In developing countries, research ethics guidelines, when they exist, generally afford less protection to participants [14,15].

those challenges, and their *ethical beliefs* about what should be done in response to those challenges.

Survey items were based on the results of 20 interviews with scholars in information technology, information systems, information studies, communication, business, and computer science.[2] All participants were faculty at U.S. and European academic institutions, or researchers in consulting or industrial research labs. The interviews asked researchers about ethical challenges they faced and how they dealt with those challenges. Qualitative coding of interviews helped us refine a list of relevant data types, existing ethical challenges, and practices to deal with those challenges. We created survey items based on this data. We also incorporated questions based on practices recommended by the formalized guidance of our institution's IRB, as well as the recommendations of the Association of Internet Researchers' (AoIR) Ethics Working Committee [24].

Survey participants first provided basic background information, followed by an assessment of online data collection and analysis practices. Participants were asked about specific challenges they faced (drawn from the qualitative study) and their personal beliefs about research ethics. At the end of the survey, researchers who were interested in receiving the project results or participating in follow-up studies could fill out an optional web form to provide an email address. The survey remained open for a period of four weeks during April and May 2015.

See Appendix for survey items and response frequencies.

Sample

The population of interest in this study is researchers who work with online user data. Therefore, we employed purposive sampling [31] to identify individuals who fit three criteria: age (18+); current employment status (i.e., doctoral student, postdoc, research scientist, faculty member, industry researcher, or otherwise in a field/organization/position that involves research with online data); and self-identifying as conducting research with online user data.

We identified eight conferences where research using online data is common and, when possible, authors come from multiple disciplines: CSCW, CHI, ICWSM, iConference, WWW, Ubicomp, CKIM, and KDD. For all conferences besides ICWSM, we compiled a list of authors on papers published since 2011 that included "trace ethnography," "big data," "twitter," "forums," "text mining," "logs," "activity traces," and/or "social network." For ICWSM, we contacted the full program committee from the previous two conferences. This resulted in approximately 2800 unique names. These participants received emails with custom links plus one reminder.

The direct email component was complemented by distribution of the survey link via social media and mailing lists targeting researchers in AoIR, AIS, CITASA, AIS ICA, STS, and NCA. This strategy increased the pool of potential researchers beyond those submitting to the identified conferences.

This sampling strategy produced a sample encompassing a wide range of researchers and research practices. Our email language and survey questions, however, focused on users of online data. We expect that many included in the original sample did not identify as "online data" researchers, and we acknowledge a self-selection bias for participants who are heavily engaged in more traditionally conceptualized "big

Variable	**Mean (SD)/N (%)**
Sex	
Male	159 (60.5%)
Female	93 (35.4%)
Education	
Bachelor's	15 (5.7%)
Master's	61 (23.2%)
PhD	180 (68.4%)
Current Location	
United States	162 (61.6%)
UK	21 (8.0%)
Canada	14 (5.3%)
Germany	12 (4.6%)
Australia	9 (3.4%)
22 other countries (<5 participants)	38 (14.4%)
Degree In…	
Business & Law	7 (2.7%)
Communication & Media	33 (12.5%)
Computer Science/Engineering	100 (38%)
Fine Arts & Humanities	9 (3.4%)
HCI	10 (3.8%)
Information	50 (19%)
Social Sciences	37 (14.1%)
Other (Bio Sciences, Education, Enviro Sciences, Physics)	11 (4.2%)
Current Field of Work	
Academia (Research Focus)	195 (74.1%)
Academia (Teaching Focus)	12 (4.6%)
Industry	35 (13.3%)
Policy/Government	7 (2.7%)
Non-Profit	5 (1.9%)
Data Activities	
Data Sources (range 0-14)	8.86 (3.33)
Research Methods (range 0-11)	7.33 (2.48)
Data Analyses (range 0-15)	9.97 (3.48)

Table 1. Sample Demographics (N=263)

[2] Data from the interviews is not presented in this paper, except as a way to provide justification for methodological decisions. Results from this work are available in [citation anonymized].

data" research. A broader discussion of the sample biases is included in the limitations section below.

Variables Included in Analyses

Data Practices

Participants were asked to list the frequency with which they engaged in three categories of research practices on a five-point Likert-type scale (1=Never, 5=Very Often): (1) 14 data sources, including online forums, tweets, activity data, location data, attitudes; (2) 11 data collection techniques, including scraping, surveys, interviews, and network data; and 14 types of analyses, including SNA, regression, path analysis, machine learning, and interpretive coding. All data practices were drawn from interview data and review of the literature. Data practices largely correlated with discipline; for example, computer scientists were more likely to use predictive analytics, data mining, data visualization, and machine learning techniques. They were also less likely to use surveys, interviews, and ethnography when compared to researchers in the social and information sciences. Finally, participants were asked to select the data sources that matter most in their research (up to three sources). The top five selections were attitudinal data (43%), activity data (37%), social network data (28%), demographics (25%), and tweets (25%).

Degree Field

Pre-survey interview data suggested that researchers believe discipline to be an important factor in attitudes towards research ethics, with several variations of statements like:

> *Oftentimes, people have very different views [on ethics], such as management, IS, communication, human computer interaction, journalism, computer science, the list goes on and on. Each have their own community and each have their own body of literature.*

Often, discussion of disciplinary norms was more pointed:

> *I think a lot of it has to do with what your background is, what disciplines you're coming out of as to whether or not you think that's a problem, because I think a lot of CS people see no problem with scraping public data anywhere, but I do.*

We therefore tracked disciplinary background in order to determine whether these assumptions— sometimes verging on accusations—held up. Participants were provided with a free text response box to list the program from which they received their degree. These responses were then categorized based on disciplines (e.g., communication and media, business, engineering), then further collapsed into overarching fields. Of these, three groups had a large enough N for meaningful comparisons: computer scientists (N=100; this group includes CS, Computer Engineering and related fields), social scientists (N=70, including communication, media studies, psychology, and sociology), and information scientists (N=62, including those from iSchools as well as those who listed "informatics").

Work Field

Another theme that arose in the pre-survey interview data was possible differences between academic and industrial research ethics practices. As one respondent put it:

> *…on the one side, you've got academic researchers being driven towards compliance with really extreme IRB. You've got the commercial guys with no IRB.*

We asked respondents to select a field of work from: (1) Academia: Research University, (2) Academia: Teaching University, (3) Industry, (4) Government/Policy, and (5) Non-Profit. We also provided an open text field and merged responses, when possible, into the relevant categories.

Attitudes Toward Acceptability of Research Practices

The survey next asked respondents about their agreement with a series of statements prefaced by "I think it's permissible to…" and "I think researchers **should**…" on a five-point Likert-type scale (1=Strongly Disagree, 5=Strongly Agree). For the prompt "I think it's permissible for researchers to…" sample items included: "Scrape data from online forums (*M*=3.98, *SD*=.89); Deceive subjects as part of research (*M*=2.61, *SD*=1.23); and Collect sensitive information from online sources (*M*=3.05, *SD*=1.16)." For the prompt "I think researchers should…" sample items included: "Ignore a website's Terms of Service when necessary to collect data (*M*=2.19, *SD*=1.21); Share research results with research subjects (*M*=3.90, *SD*=.81); and Remove individuals from datasets upon their request (*M*=4.55, *SD*=.71)."

Attitudes Toward Other Researchers' Ethics

Based on researchers' perceptions of disciplinary differences reported in the interviews, a final series of attitudinal items gauged respondents' perceptions of how ethical codes vary across disciplines or work sites based on agreement with provided statements (1=Strongly Disagree, 5=Strongly Agree). The three most relevant questions were: (1) Researchers are held to a higher ethical standard than others who use online data (*M*=3.88, *SD*=.92); (2) Industry researchers are held to a higher ethical standard than academic researchers who use online data (M=2.20, *SD*=1.04); and (3) Academics consider the ethical implications of their online data collection more than industry researchers (*M*=3.56, *SD*=1.11).

Personal Code of Ethics

Finally, we prompted participants with open-ended questions to gain insights into the strategies they employed to ensure the quality and security of online data. In this analysis, we include a discussion of responses to the question: "How would you describe your personal code of ethics regarding online data?" We received 162 (63%) responses, with an average length of 35 words (median=23.5; *SD*=35).

Data Analysis

We downloaded and imported participant data into SPSS for analysis. We cleaned the data by removing cases missing more than 10% of responses. As the presentation of findings is largely descriptive, we did not impute missing data; these cases were ignored in individual analyses.

For the open-ended question, each author independently reviewed the responses and created a set of codes to apply to the corpus. After three iterative rounds of comparison, authors agreed on a 11-factor coding scheme (see Table 2). Two authors then coded the full set of responses for the 11 themes—allowing for multiple codes per response—and agreed on 97% of the codes. For the final 3%, we discussed areas of disagreement until we reached a consensus. Key findings from these analyses are presented below.

FINDINGS

We present findings in four areas: describing how researchers characterize their codes of ethics; identifying areas of agreement and disagreement among respondents; analyzing researchers' impressions of their *colleagues'* ethical standards; and identifying an emergent code of ethical attitudes among our respondents.

Researchers' Codes of Ethics

We combined structured and semi-structured survey questions to understand respondents' personal codes of ethics for research with online data. The most frequent responses to the open-ended question asking participants to share their personal codes of ethics described strategies for *protecting individuals*, *securing consent*, and *balancing risks to participants with larger (i.e., societal) benefits*. Respondents also cited diverse ethical principles, including the Belmont Report and Common Rule, guidance from ethics review boards, individual sites' Terms of Service, the Hippocratic Oath, and variations on the Golden Rule. The role of context in guiding researcher decisions was also a prominent theme, as were changing personal codes of ethics. We explore descriptive statistics and qualitative data highlighting these themes below.

Protecting data subjects

Protection of individual subjects was the most prominent theme in free-text responses, and responses took many forms. A male research faculty member working in an English department characterized protection as de-identification:

> *Every effort should be taken to present any results from research using online data in a manner which does not allow for the identification of individuals whose actions may be recorded in the data.*

Researchers reported achieving de-identification by reporting only aggregated data, or by not collecting,

Code	Definition	Example Statements
Public Data	Only using public data / public data being okay to collect and analyze	*In general, I feel that what is posted online is a matter of the public record, though every case needs to be looked at individually in order to evaluate the ethical risks.*
Do No Harm	Comments related to the Golden Rule	*Golden rule, do to others what you would have them do to you.*
Informed Consent	Always get informed consent / stressing importance of informed consent	*I think at this point for any new study I started using online data, I would try to get informed consent when collecting identifiable information (e.g. usernames).*
Greater Good	Data collection should have a social benefit	*The work I do should address larger social challenges, and not just offer incremental improvements for companies to deploy.*
Established Guidelines	Including Belmont Report, IRBs Terms of Service, legal frameworks, community norms	*I generally follow the ethical guidelines for human subjects research as reflected in the Belmont Report and codified in 45.CFR.46 when collecting online data.*
Risks vs. Benefits	Discussion of weighing potential harms and benefits or gains	*I think I focus on potential harm, and all the ethical procedures I put in place work towards minimizing potential harm.*
Protect Participants	Methods to protect individual: data aggregation, deleting PII, anonymizing/obfuscating data	*I aggregate unique cases into larger categories rather than removing them from the data set.*
Deception	Justifying its (non) use in research	*I use deception for participatory research and debrief at the end.*
Data Judgments	Efforts to not make inferences or judge participants or data	*Do not expose users to the outside world by inferring features that they have not personally disclosed.*
Transparency	Contact with participants or methods of informing participants about research	*I generally choose not to scrape/crawl public sources. I prefer to engage individual participants in the data collection process, and to provide them with explicit information about data collection practices.*
In Flux	One's code of ethics is under development, context-dependent, or otherwise in flux	*It very much depends on the nature of the data.*

Table 2. Emergent themes from qualitative responses regarding researchers' personal code of ethics

obscuring, or replacing names or other identifying information. For some, obscuring identifying information extended to altering data. As one academic from IS wrote:

> *Direct quotes from text posted online should be used with permission of the poster or obfuscated to reduce 'googleability'.*

This comment reflects an unintended consequence of data's persistence online. While beneficial for many types of research, persistence also makes it more difficult for researchers to protect individuals' identity when analyzing and sharing public data.

Data sharing

Half of the respondents reported making datasets available to other academics (through restricted release) or the public. However, some respondents indicated they avoided sharing data with others to prevent re-identification, highlighting tensions between advancing knowledge and protecting individuals. For example, a male IS professor said:

> *I do not share raw datasets. I do not discuss individuals or unique participants. I inspect the results before reporting with an eye towards identifying any potential privacy breaches or other negative consequences.*

Others said they distribute data under agreements specifying use for "research and development" purposes only. One researcher in a communication department indicated he uses a verification process before sharing data with other researchers:

> *In the case of public, but still potentially risky data, we've required that individuals provide names, titles, etc. so that we can verify that they are researchers…*

Responses like this suggest that researchers believe they can trust other researchers to use shared data responsibly.

In addition to data sharing, we asked respondents about sharing practices for code they create to scrape online social media data (which can be instrumental in attaining online data) as well as strategies for protecting that code. Only 29% of respondents reported making code available. One IS respondent described her heuristics around sharing code:

> *I don't release code that violates TOS (e.g. scrapers) but I do release code that uses APIs.*

Securing informed consent

Many researchers reported valuing consent of participants, although the operationalization of consent varied significantly across participants. Only 22% of respondents agreed that informed consent is "always" necessary for online data collection, while the percentage rises to 52% for studies collecting "personal information," 44% for studies that collect data from minors, and 65% for collecting "potentially sensitive information." Behaviors about consent mirrored beliefs: 35.6% of respondents reported never having used online data without consent while another 34.8% reported using online data without consent often or very often. Qualitative responses reflected this division in practice. For example, a media studies researcher wrote:

> *As long as the population agrees with the goals of your study and you are transparent, you are good to go…*

An HCI researcher responded:

> *My experiences have been to ... gain consent first by viewing the content with the owner…*

An IS researcher put it this way:

> *Community members should be offered the opportunity to have their identity associated with their quotes/content if they so desire and if it is appropriate for the research context.*

Many researchers went beyond consent in descriptions of their ethical codes of conduct, describing high standards for transparency. For example, 30.5% of respondents reported sharing research results with subjects often or very often, and another 33.6% reported doing so "a few times." Free-text responses also described ethical codes for transparency. As one computer scientist wrote:

> *Inform your participants! (And do it better than the standard IRB forms)…*

An anthropologist wrote:

> *I also try to communicate openly with my research subjects what we do, why we do it and what they give and gain.*

A communication professor put it this way:

> *I try as hard as possible to be clear on how and why I'm collecting/working with data whenever I'm harvesting, as well as open to participant's opinions during or after the collection.*

Minimizing, balancing, or doing no harm

Many respondents offered some variation on the Hippocratic Oath when asked to state their personal code of research ethics: several wrote "Do no harm" as their primary or only response. In survey responses, 56.9% of respondents agreed or strongly agreed with the statement "I think researchers should only collect online data when the benefits outweigh the potential harms." Researchers who focused their personal codes of ethics on preventing harm to participants varied from trying to *minimize* harm ("Make sure there is minimal risk to the participants") to trying to prevent harm entirely ("never hurt participants from any aspects.") Most researchers writing about harm seemed to trust their own judgment about what might prove harmful. A few, however, qualified their ability to foresee harm. For example, a sociologist working in industry clarified: "Do no intentional harm."

Beyond the Belmont Principles

Responses focused on protecting individuals, securing informed consent, and balancing research risk and harm emphasize that the Belmont Principles still influence social computing researchers' conception of ethical research practice. As one IS respondent said:

> *I focus on the principles in the Belmont Report and try to stick to those.*

This was echoed in many responses, reflecting that researchers actively consult with their IRBs in order to conduct responsible research.[3] Sample responses included this one from an IS researcher:

> *I don't have a personal code of ethics. I try to follow the ethics in the Common Rule and IRB/CPHS standards.*

The Common Rule/Belmont Principles were not the only principles cited by researchers. Variations on the Golden Rule also provided a common framework for respondents. A few respondents cited the Golden Rule directly by name. Others, such as this psychology researcher, paraphrased:

> *In other words: do not do to others what you don't want others to do upon you.*

Likewise, researchers empathized with research subjects in statements such as this, from an academic in IS:

> *[I] ask myself, "would I participate in this study?"*

We also received responses from many respondents who felt that Belmont Principles, particularly informed consent, were less relevant because they were collecting *public* data. However, there was disagreement among respondents about what constitutes public data. Researchers had diverse ways of defining whether online data was also public data; for example, one IS researcher said:

> *If the data is available on the internet without logging in, it is permissible…*

Another researcher in Computer and Information Sciences expressed the similar sentiment:

> *If the data is volunteered without any expectation of privacy (as on Twitter) it is ok to use that data for aggregate analysis…*

An HCI professor noted that IRBs support this assertion:

> *I focus almost exclusively on Twitter data which multiple IRBs have concluded is public domain information.*

Other researchers disagreed with this evaluation of public data, however. A CS professor wrote:

> *I try to consider whether the research context is a significant departure from the original context the data were published in, before embarking on collection. For this reason, I generally choose not to scrape/crawl public sources.*

Almost all researchers who cited the public nature of online data also said they try to anonymize or avoid identifying individuals in public datasets, signaling that even "public" data is still seen as sensitive by most.

Agreement and Disagreement in Ethical Attitudes

We evaluated the variation in responses to 30 Agree/Disagree statements about research attitudes and practices to establish where our sample found common ground and where they expressed significant differences. We grouped responses to these items based on two criteria: mean score and mean variance. We then created four categories based on level of agreement (with responses below the midpoint categorized as "disagreement" and responses above 3.5 classified as "agreement") and variance (with variance scores below 0.8 classified as "low variance" and above 1.2 as "high variance"). As seen in Table 3, no items met the criteria for disagreement/low variance, i.e., there were no statements with which a majority of participants disagreed.

Five items were cohesive across respondents, suggesting a set of foundational research practices for conducting research using online user data. These included removing a subject from a dataset when the individual formally requests it; talking to colleagues and review boards about ethical considerations in one's research; making research

	Low variance (<.8)	**High variance (> 1.2)**
Agreement (>3.5)	• Remove subjects from datasets upon request[1] • Ask (1) colleagues or (2) IRBs about research[1] • Share results with participants[1] • Think about edge cases/outliers[1]	• Use non-representative samples[2] • Remove unique individuals before sharing[1] • Researchers can't collect large-scale online data if consent is required
Disagreement (<3)	**No items.** *Notes:* [1] "I think researchers should…" [2] "It's permissible for researchers to…"	• Ignore ToS when necessary[2] • Deceive participants[2] • Share raw data with key stakeholders[1] • It's possible to obtain informed consent with large-scale studies

Table 3. Comparing level of agreement and item variable across the corpus (N=263).

[3] 58.2% of respondents said they have discussed research ethics with their IRB or internal research board at least once, with 22.4% responding that they do this often or very often.

results (not raw data) available to participants upon study completion; and being careful about reporting on edge cases and outliers. When looking at the corpus of responses with significant variance (i.e., greater than 1.2), we expected these differences to be attributed to disciplinary backgrounds; however, only deception differed significantly across disciplines, with CS and IS scholars expressing significantly lower agreement than communication scholars (but not all social scientists). Deception plays a significant role in many aspects of communication research [18], but the term may be more controversial for researchers from other traditions. A separate question in the survey asked participants to describe situations when deceiving participants was acceptable, and while only a minority (21%) said never, most respondents noted that informed consent had to be obtained and/or the deception had to pose a minimal risk to participants. Likewise, individuals generally agreed that researchers should remove unique cases from datasets before sharing more widely. While this seems reasonable—as the Taste, Ties, and Time study [21] illustrated, re-identification of unique individuals requires little technical skill)—the high variance suggests that a subset of researchers (in this case, computer scientists) have significantly more negative attitudes toward practices that might compromise data integrity.

Impressions of Colleagues' Ethical Standards

In this section, we evaluate how attitudes vary for three statements that framed different groups—academics, industry researchers, and researchers in general—as "more ethical." This line of questioning arose from a pervasive suggestion in earlier interview data that researchers, and particularly academic researchers, were being held to an unfair ethical standard by the public and/or by IRBs. We worded corresponding survey items to ask participants whether they believe specific types of researchers are held to higher ethical standards than others, and to observe how individual characteristics may influence these attitudes. We conducted a series of between-subjects ANOVAs (see Table 4) with *Tukey's B* post-hoc tests to determine if different types of researchers held different attitudes about the ethics of others in the field, and to observe if subsets of groups (based on degree and job) differed. All three ANOVAs were significant, and differences emerged for all items within degree program and for one item within field of work. No differences emerged based on location, although this analysis was limited to a dichotomous comparison because the sample size for groups outside the U.S. were too small and all attempts to categorize the remaining countries produced highly variable results, suggesting that norms are not consistent by region.

Looking at the differences in responses across disciplines, some noteworthy themes emerge. Social scientists largely believe that academics consider the ethical implications of their work more deeply, while computer scientists suspect there is less difference. Information scientists' impressions of their colleagues' standards are not quite in line with the social scientists but are clearly different than computer scientists. Unsurprisingly, industry researchers as a whole see the ethical standards of academic and industry research as similar, while academics as a whole disagree. CS researchers may be less judgmental about ethical standards in industry research due to cross-fertilization: they may be more likely to have interned or worked in industry or collaborated with former students in industry.

Codification of Ethical Attitudes

Our final analysis evaluates whether individual characteristics are associated with a more codified set of ethical beliefs, attitudes and practices by reporting agreement with items more closely aligned with formal

	Academics consider ethical implications more	Industry researchers are held to a higher ethical standard	Researchers are held to a higher ethical standard
Degree Program			
Social Science	3.94^{b}	1.91^{a}	4.21^{b}
Computer Science	3.36^{a}	2.41^{b}	3.68^{a}
Information Science	3.49^{a}	2.03^{a}	4.02^{b}
Field of Work			
Academia	3.69^{b}	2.07^{a}	3.93^{a}
Industry	2.97^{a}	2.69^{a}	3.90^{a}
Government/Nonprofit	3.42^{ab}	2.23^{a}	4.08^{a}
Current Location[1]			
US	3.53	2.12	3.91
Other	3.64	2.28	3.83
Test Between Subjects Effects	F(6, 215)=4.33 p<.001	F(6, 215)=5.13) p<.001	F(6, 214)=2.71 p=.015
Adjusted R^2	.083	.101	.044

Notes: [1] *Sample size differences prevented analysis at a more granular level than US v. Everyone Else, which likely plays a role in non-significant findings.* ^{ab} *Superscript letters show groupings based on Tukey's B post-hoc tests.*

Table 4. Between-subjects ANOVA on attitudes researcher ethics.

Item	*M*	*SD*
...notify participants about why they're collecting online data[1]	3.89	0.96
...share research results with research subjects[1]	3.90	0.80
...Ask colleagues about their research ethics practices[1]	4.27	0.74
...Ask their IRB/internal reviews for advice about research ethics[1]	4.03	0.90
...Think about possible edge cases/outliers when designing studies[1]	4.33	0.71
...Only collect online data when the benefits outweigh the potential harms[1]	3.62	1.10
...Remove individuals from datasets upon their request[1]	4.56	0.71
Researchers *should* be held to a higher ethical standard than others who use online data[2]	3.46	1.22
I think about ethics a lot when I'm designing a new research project[2]	3.96	0.93

[1] Prompt: "I think researchers should...." [2] Prompt: "To what extent do you agree with the following statements?"
Both sets of items were measured on five point, Likert-type scales (Strongly Agree-Strongly Disagree).

Table 5. Codification of Ethical Attitudes Measure

ethical codes such as the Belmont Report or rules specified by ethics review boards. In the survey, we asked participants a series of questions about their attitudes toward, and engagement with, various research practices (see Appendix). Through exploratory factor analysis of 35 items, we created a reliable nine-item measure (α=.71, *M*=4.00, *SD*=.49) that captures attitudes toward a variety of behaviors drawn from IRB codes of conduct, AoIR recommendations, and earlier interview data about emerging online research practices. Participants with a higher score on this scale also report spending more time reflecting on and talking with others about ethical aspects of their research. We characterize respondents who agree with items in this scale to have a more *codified* set of ethics practices (see Table 5).

To identify if a more codified set of ethics practices was associated with individual characteristics, we conducted an OLS regression (Table 6). We created dummy variables for academic degree (social, computer, information science) and field of work (academia, industry, policy/non-profit). The analysis yielded only one significant predictor—work field—with *Tukey B* post-hoc analyses finding that industry workers' average agreement (*M*=3.75) was significantly lower than non-profit/policy workers (*M*=4.11). Academics (*M*=4.02) were not significantly different from either group. No other variables included were significant and the amount of variance explained by this model was minimal.

DISCUSSION

In many ways, the increase in "big data analytics" in the 21st century is similar to earlier developments in offline data collection, such as Nielson's market analyses for radio and television in the first half of the 20th century. As the quantity and accessibility of content about people grows, so do questions about how to conduct appropriate and ethical research.

Independent Variables	*ß* (t-test)	*p*
Sex: Female	.01 (.13)	.90
Academic Degree		
Social Science (1)	-.06 (-.72)	.48
Computer Science (1)	-.13 (-1.45)	.15
Field of Work		
Academia (1)	-.16 (-1.23)	.22
Industry (1)	**-.31 (-2.43)**	**.02**
Level of Education	.01 (.07)	.94
Location: U.S.	-.01 (-.20)	.84
F-test	1.56	.15
Adjusted R^2	.02	--

Table 6. OLS Regression for Codified Ethics

The present study offers new insights into the attitudes and practices of social computing researchers working with online data. A primary theme emerging from our qualitative analysis was the dependency between ethical principles. Many researchers discussed using public data without consent, but taking precautions to de-identify individuals when they did so. Likewise, researchers collecting sensitive or high-risk information largely felt this practice *demanded* informed consent. These dependencies highlight the contextual nature of research ethics and may help explain the lack of significant findings in the regression analysis. As one IS researcher put it:

> *I'm not convinced that my personal approach is fully appropriate for other researchers and datasets.*

Other researchers expressed that navigating contextual complexities was best left to trained researchers, as they should be trusted to determine the best solution for their particular confluence of methods, data types, and topics.

The data also indicate that neither discipline nor use of particular research methods correlate to differences in research ethics practices with a single exception.[4] Differences in ethical beliefs are therefore not primarily attributable to discipline or method (as widely suspected in earlier interviews), or to gender or cultural factors (here captured as geographic location). Differences may be related to variables not measured here, such as personal

[4] Among participants who collect data using third-party applications, increases in the frequency of this practice were associated with decreases in agreement with codified ethics measure Note that there were no significant differences when treating this data collection practice as a dichotomous variable (i.e., never vs. ever).

attributes or experiences, or norms in particular research groups. Follow-up research is needed to further examine reasons for variations in researchers' beliefs.

That said, the data point to a disciplinary divide in *impressions* of the ethical standards in non-academic research domains. These perceived differences, however, are not reflected in practice; the difference between industry and academic agreement with the 'codified' measures is not significant. This finding suggests that, if we focus on disciplinary differences, or differences between academic and industry research, *we risk talking past each other* rather than learning from others' experiences. After all, some of the most creative work in research ethics review is taking place in industry research settings [4]. Discussions of research ethics should be rooted in practice if we are to have a constructive debate around how to update ethical principles for online research.

Developing Ethics Heuristics for Online Data Research

Findings from this study highlight a number of areas where researchers largely agree on what constitutes ethical research. We draw upon these findings to propose heuristics for conducting ethical research that move beyond the Belmont Report. Participant responses on the codification of ethical attitudes measure can be grouped into four categories. One attitude echoes the current guidance of the Belmont Report: researchers with a codified set of ethical attitudes believe they should only collect online data when the benefits outweigh the potential harms. But there are three categories of emerging beliefs and practices that go beyond the Belmont Report's recommendations: (1) transparency with participants, (2) ethical deliberation with colleagues, and (3) caution in sharing results.

Transparency with research communities is an important part of ethical practice for online research. Agreement with statements that researchers should "notify participants about why they're collecting online data," (66% agreement) "share research results with research subjects," (69.2% agreement) and "remove individuals from datasets upon their request" (91.5% agreement) all highlight the importance of transparency in online data research. These practices require either a consent mechanism or a degree of transparency with data subjects.

Transparency entails a range of practices, from notification before data collection to debriefing after, and can take many contextually-appropriate forms. We suggest that transparency focus both on *intent* (what you are doing with data and why) and *practice* (how you're getting the data). Transparency is a flexible principle that enables subjects to both understand their participation in research and request removal from datasets if necessary. Achieving transparency, however, may be more difficult for some kinds of data collection (e.g., large-scale collection of Tweets) or for data analyzed by platform hosts. Creativity in modes of transparency is an open area for research ethics innovation, and will ideally involve collaboration across disciplines and work environments.

Ethical deliberation with colleagues in addition to ethics review boards is another important part of ethical practice for online research that goes beyond the Belmont principles. In the codification of ethical attitudes measure, this is captured in agreement with statements that researchers should ask colleagues about their research ethics practices (87.1% agree), and ask their IRB/internal reviewers for advice about research ethics (73.3% agree). This principle maps to AoIR's broader emphasis on a deliberative process, including to "consult as many people and resources as possible" [24]. We agree with this best practice and emphasize that expanding the pool of resources *beyond* direct colleagues is an important step for researchers. Colleagues may struggle to be honest in their assessment of projects; relative isolation from social pressures is an advantage of review bodies such as IRBs. Researchers should discuss projects with review boards before performing any online data collections. Even if not strictly required by current IRB standards, such discussion will help both researchers and review boards to clarify best practices and enhance the review process for future projects. In turn, we are hopeful these discussions will help review boards better understand changing technological research practices, and become better resources for evaluating online research ethics.

Finally, the codification of ethical attitudes measure suggests that researchers should be *cautious about sharing results that include (potentially identifiable) outliers*, with 88.6% of respondents agreeing with this principle. However, such guidance does not specify what constitutes "careful." Best practice for taking care with outliers is hard to define and likely varies on a case-by-case basis. We believe researchers can address ethical concerns surrounding outliers (e.g., their identifiability within a dataset) by seeking outside advice and feedback as part of a deliberative ethical process. Ethical considerations for *reporting* data highlights AoIR's guidance that ethical challenges can occur throughout the research process [24], and that researchers should consider consulting with colleagues and review boards at later points in the research process than is traditional. For example, issues with the release of the T3 dataset [21] might have been avoided had outside parties with deeper knowledge of the site and anonymization pitfalls been consulted [38].

Our paper also points to areas of significant disagreement in the online data research community: use of non-representative samples; removal of unique individuals from datasets; the tension between obtaining consent and collecting data from some sources (and whether it is possible to obtain informed consent for large scale studies at all); the ethics of ignoring Terms of Service; the ethics of deceiving participants; and the necessity of sharing data with research subjects. These are critical areas of disagreement on which to focus consensus-building efforts.

That said, ethics is not only a process of consensus-building around best practices; ethical principles are not made by majority rule. Researchers may disagree on practices that ethicists, policymakers, or the public feel are important. Further, context-dependent factors will prevent full consensus on *all* practices. We have therefore drawn upon ongoing work in research ethics to reflect on areas of disagreement in our data, as well as issues not reflected in our data. Our final recommendations focus on bias, consent, and contextual norms.

First, we believe researchers should report on the makeup of their online samples, and particularly potential biases in their samples. This concern has been well-documented by researchers [12] and we wish to emphasize it here.

Second, we recognize that gaining informed consent is not a reasonable requirement for all open online datasets. But we would caution researchers to—whenever possible—respect the norms of the contexts in which online data was generated. Data are generated with particular information flows in mind, and users are often very aware of those flows [28]. Twitter has different information norms and expectations than Reddit threads or software development forums. Researchers have an obligation to try to understand these norms, as well as if those norms dictate particular transparency or consent mechanisms.

A focus on contextual norms also underscores AoIR's guiding principle that *harm* is contextually defined [24]. We believe survey respondents' emphasis on versions of the Golden Rule—do not create a study you wouldn't want to participate in—provides an excellent starting benchmark for evaluating potential harms. But the Golden Rule is not a *sufficient* principle. Researchers should also consider potential consequences—and potential benefits—for the most vulnerable members of a population before beginning data collection.

Our findings highlight that there is not yet consensus on the ethics of breaking Terms of Service (e.g., regarding data scraping) and the proper role of deception in online studies. Indeed, the three authors of this paper have differing opinions on these issues. We hope that highlighting these points of contention will spur further studies to guide decision-making when engaging in this kind of research.

Finally, our study illustrates that researchers in a variety of social computing fields are thinking deeply about research ethics in their work. Perhaps as a result of recent media attention to research ethics, or because ongoing educational efforts, it is clear that considering responsible conduct of research is a part of many researchers' practice.

We believe this increased attention on ethical research is positive and beneficial for the research community at large and CSCW in particular. The CSCW community stands at a critical moment in the study of ethical research practices in online data collection. We must not shy away from tough discussions about responsible research practices but rather embrace the challenges associated with conducting meaningful and ethical social computing research.

Risks and Limitations

Determining the total population of international researchers performing research using online data is a difficult task, making sampling especially challenging. While all attempts were made to reach as many researchers working with online data as possible, we may have inadvertently excluded population subsets due to the diversity of researchers, topics, and venues.

It is possible that participants who chose to take the survey were more interested or expert in research ethics; that said, open-ended responses indicated a wide range of experiences with research ethics concerns. Responses such as "I have given some thought [to a personal code of ethics] but probably not enough" indicate this diversity.

Because this survey asked about controversial topics, there is a risk of social desirability bias in participant responses. All survey responses were anonymous to reduce the risk of bias (which we reminded participants throughout the survey), and questions were worded as neutrally as possible.

Survey questions were also not designed to elicit *contextual* variables, which some respondents (rightly) pointed out might impact responses. It is important to note that many researchers' beliefs about the ethics of particular practices might change based on the purpose or context the study being conducted. Delineating all possible contextual variables is a difficult challenge for early-stage research targeted at a broad community that uses many types of methods for many purposes. Future work might target narrower research communities in order to use context-sensitive survey methods (e.g. [16, 32]) to better understand the influence of contextual variables on ethical beliefs and judgments.

CONCLUSION

Online spaces that focus on content sharing and user connection have opened up a wealth of new research opportunities for those who study human relationships, technology adoption, user experience, algorithmic work, and much more. Researchers in these areas express beliefs and engage in practices that demonstrate that the guiding ethical principles for research encapsulated in the Belmont Report nearly 40 years ago no longer provide sufficient guidance for research conducted with large and diverse sets of online data. The challenges of online data collection are extraordinarily nuanced, and reflect problems based on the difficulties of defining contexts and norms in online spaces. While respect for persons, beneficence, and justice remain meaningful values to which to aspire, the ways that these principles have been interpreted are not specific enough for online research.

This paper describes codified principles to which many social computing researchers report adhering in belief and practice. It also provides data that disprove narratives of difference in either disciplinary or industry/academic divisions. Taken together, the study provides evidence for interdisciplinary research communities such as CSCW to approach the challenge of building research frameworks that both encourage new methods and analyses of online data and ensure that such research is conducted responsibly. We believe that studies like this one represent an important opportunity to engage the research community in designing ethical guidelines.

ACKNOWLEDGEMENTS

Blank for review.

REFERENCES

1. June Ahn. 2012. Teenagers and social network sites: do off-line inequalities predict their online social networks? *First Monday* 17, 1.
2. Eytan Bakshy, Solomon Messing, and Lada Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science,* aaa1160.
3. Michael Barbaro and Tom Zeller. 2006. A Face Is Exposed for AOL Searcher No. 4417749. *The New York Times*. Retrieved from http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all
4. Anne Bowser and Janice Y. Tsai. 2015. Supporting Ethical Web Research: A New Research Ethics Review. *Proceedings of the 24th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 151–161.
5. danah boyd. 2011. White flight in networked publics? How race and class shaped American teen engagement with MySpace and Facebook. In *Race After the Internet*, Lisa Nakamura and Peter Chow-White (eds.). Routledge.
6. danah boyd and Kate Crawford. 2012. Critical questions for big data. *Information, Communication & Society* 15, 5, 662–679.
7. Rafael Capurro and Christoph Pingel. 2002. Ethical Issues of Online Communication Research. E*thics and Information Technology* 4, 3, 189–194.
8. Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting Postpartum Changes in Emotion and Behavior via Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* ACM, 3267–3276.
9. Casey Fiesler, Alyson Young, Tamara Peyton, et al. 2015. Ethics for Studying Online Sociotechnical Systems in a Big Data World. *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, ACM, 289–292.
10. Danyel Fisher, David W. Mcdonald, Andrew L. Brooks, and Elizabeth F. Churchill. 2010. Terms of Service, Ethics, and Bias: Tapping the Social Web for CSCW Research. *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, ACM.
11. Vindu Goel. 2014. As Data Overflows Online, Researchers Grapple With Ethics. *The New York Times*. Retrieved April 8, 2015 from http://www.nytimes.com/2014/08/13/technology/the-boon-of-online-data-puts-social-science-in-a-quandary.html
12. Eszter Hargittai. 2015. Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science* 659, 1, 63–76.
13. Luke Hutton and Tristan Henderson. 2013. An Architecture for Ethical and Privacy-sensitive Social Network Experiments. *SIGMETRICS Perform. Eval. Rev*. 40, 4, 90–95.
14. A. A. Hyder, S. A. Wali, A. N. Khan, N. B. Teoh, N. E. Kass, and L. Dawson. 2004. Ethical review of health research: a perspective from developing country researchers. *Journal of Medical Ethics* 30, 1, 68–72.
15. Adnan A. Hyder, Bridget Pratt, Joseph Ali, Nancy Kass, and Nelson Sewankambo. 2014. The ethics of health systems research in low- and middle-income countries: A call to action. *Global Public Health* 9, 9, 1008–1022.
16. Guillermina Jasso. 2006. Factorial Survey Methods for Studying Beliefs and Judgments. *Sociological Methods & Research* 34, 3, 334–423.
17. Rob Kitchin. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications Ltd, Thousand Oaks, CA.
18. Mark L. Knapp, Roderick P. Hart, and Harry S. Dennis. 1974. An Exploration of Deception as a Communication Construct. *Human Communication Research* 1, 1, 15–29.
19. Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24, 8788–8790.
20. Robert Kraut, Judith Olson, Mahzarin Banaji, Amy Bruckman, Jeffrey Cohen, and Mick Couper. 2004. Psychological research online: report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist* 59, 2, 105.

21. Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. 2008. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks* 30, 4, 330–342.
22. Mary Madden. 2012. Privacy management on social media sites. Washington, D.C. Retrieved May 9, 2015 from http://www.pewinternet.org/2012/02/24/privacy-management-on-social-media-sites/
23. Bradley Malin and Latanya Sweeney. 2004. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedial Informatics* 37, 3, 179–192.
24. Annette Markham and Elizabeth A. Buchanan. 2012. Ethical decision-making and internet research. Association of Internet Researchers. Retrieved from http://aoir.org/reports/ethics2.pdf
25. Deborah Martin. 2007. Bureacratizing Ethics: Institutional Review Boards and Participatory Research. *ACME: An International E-Journal for Critical Geographies* 6, 3, 319–328.
26. A. Narayanan and V. Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. *IEEE Symposium on Security and Privacy*, 2008. SP 2008, IEEE, 111–125.
27. Jeffrey Nichols and Jeon-Hyung Kang. 2012. Asking Questions of Targeted Strangers on Social Networks. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, ACM, 999–1002.
28. Helen Nissenbaum. 2009. *Privacy in context: technology, policy, and the integrity of social life*. Stanford Law Books, Stanford, CA.
29. Office of the Secretary of The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Department of Health, Education, and Welfare.
30. Paul Ohm. 2010. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review* 57, 1701.
31. Michael Quinn Patton. 2001. *Qualitative Research & Evaluation Methods*. SAGE Publications, Inc, Thousand Oaks, Calif.
32. Peter H Rossi and Steven L Nock. 1982. *Measuring social judgments: the factorial survey approach*. Sage Publications, Beverly Hills.
33. Janet Salmons. 2014. New Social Media, New Social Science... and New Ethical Issues! Google Docs. Retrieved May 6, 2015 from https://drive.google.com/file/d/0B1-gmLw9jo6fLTQ5X0oyeE1aRjQ/edit?usp=sharing&usp=embed_facebook
34. Christian Sandvig. The Facebook "It's Not Our Fault" Study. Social Media Collective. Retrieved May 21, 2015 from http://socialmediacollective.org/2015/05/07/the-facebook-its-not-our-fault-study/
35. Paul M. Schwartz and Daniel J. Solove. 2011. The PII Problem: Privacy and a New Concept of Personally Identifiable Information. *New York University Law Review* 86, 1814.
36. Katie Shilton. 2012. Participatory personal data: an emerging research challenge for the information sciences. *Journal for the American Society of Information Science* 63, 10, 1905–1915.
37. Zeynep Tufekci. 2015. How Facebook's Algorithm Suppresses Content Diversity (Modestly) & How the Newsfeed Rules the Clicks. Medium. Retrieved May 21, 2015 from https://medium.com/message/how-facebook-s-algorithm-suppresses-content-diversity-modestly-how-the-newsfeed-rules-the-clicks-b5f8a4bb7bab
38. Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology* 12, 4, 313–325.
39. Doug Zytko, Jessa Lingel, Jeremy Birnholtz, Nicole B. Ellison, and Jeff Hancock. 2015. Online Dating As Pandora's Box: Methodological Issues for the CSCW Community. *Proc CSCW*, ACM, 131–134.

Appendix

Below we present a subset of the survey items with the percentage of participants who responded for each category.

How often have you collected the following types of online data?	**Never**	**Rarely**	**Sometimes**	**Often**	**Very Often**
Online forum Posts	25.7%	16.4%	25.7%	20.1%	12.3%
Tweets	35.2	12.8	16.8	16.8	18.3
Dyadic interaction data	43.7	15.7	17.9	12.3	10.4
Group interaction data	33.0	13.7	28.9	12.6	11.9
Search history	62.5	21.6	9.3	4.5	2.2
Demographic data	17.2	12.8	22.0	19.4	28.6
User activity data (likes, favorites, clicks)	22.1	11.1	26.6	19.6	20.7
Social network data	22.6	12.6	22.6	23.3	18.9
Biometric data	84.3	6	5.2	3.4	1.1
Location data	36.2	15.5	23.6	16.2	8.5
Visual data	30.9	17.6	28.7	14.3	8.5
Metadata	28.0	14.9	23.1	19.0	14.9
Attitudinal data	20.3	12.5	23.6	22.9	20.7
Health data	66.2	14.7	8.5	5.1	5.5

How often do you use the following methods in online settings?	**Never**	**Rarely**	**Sometimes**	**Often**	**Very Often**
A/B testing or multivariate experimental design	43.0%	16.5%	19.9%	12.9%	7.7%
Surveys	12.2	15.5	24.0	27.3	21.0
Scraping publicly available data	13.6	16.1	29.3	19.0	22.0
Working with a site/platform/company to collect server-level data	45.1	18.2	19.3	11.3	6.2
Interviews	23.7	13.7	22.2	19.6	20.7
Ethnography	39.8	13.8	20.8	12.6	13.0
Data collection from crowdworkers	42.8	20.8	17.5	10.0	8.9
Ego/group network data collection	51.7	17.8	17.1	7.4	5.9
Data harvest using an application/extension developed by you/your research team	24.2	12.5	18.7	22.0	22.7
Data harvest using an application/ extension developed by third parties	40.2	22.9	22.5	9.6	4.8
Building algorithms/classifiers/machine learning to draw conclusions about users	36.0	13.6	19.1	15.4	15.8

How often do you use the following analytical methods on data gathered from online sources?	**Never**	**Rarely**	**Sometimes**	**Often**	**Very Often**
Data Visualization	16.0%	9.5%	26.2%	23.6%	24.4%
Social network analysis	26.3	22.6	22.6	16.8	11.3
Sentiment analysis	38.2	23.9	19.9	11.0	6.6
Other types of linguistic analysis	34.0	15.3	22.0	18.3	10.1

Descriptive statistics (e.g., means, t-tests)	10.9	7.3	19.3	26.9	35.3
Correlational analysis (including chi-square)	19.6	13.5	24.0	25.1	17.5
ANOVA/Regression	21.5	12.0	23.4	22.6	20.1
Path Models/SEM	66.3	17.8	8.9	4.8	1.9
Mediation Analysis	72.6	14.1	7.8	3.7	1.5
Interpretive analyses (e.g., grounded theory)	25.1	16.6	17.7	21.4	18.8
Longitudinal analyses	24.2	18.7	26.7	17.2	12.8
Analysis by crowdworkers (e.g., Mechanical Turk)	56.3	19.1	13.6	5.5	5.1
Predictive analytics	43.4	14.3	22.8	10.3	8.8
Data mining	33.8	15.6	18.2	16.0	16.0
Machine learning	36.3	15.6	18.1	13.0	16.7
Content analysis	13.5	9.5	27.6	29.1	20.0
How frequently have you done the following things as part of your digital data collection?	**Never**	**Rarely**	**Sometimes**	**Often**	**Very Often**
Scraped data from online forums	43.3%	8.1%	25.9%	12.2%	10.4%
Used online data without explicit participant consent	35.6	7.9	21.7	17.2	17.6
Reused data collected for one purpose to analyze a completely new research question	39.7	7.5	28.5	13.9	10.5
Used online data from a non-representative sample	32.6	6.8	27.3	15.9	17.4
Inferred characteristics of **individuals** based on online data	52.1	4.9	21.7	13.1	8.2
Violated a platforms terms of service (e.g., terms that prohibit scraping)	72.7	5.2	16.1	4.1	1.9
Used a platforms API to collect data	26.8	7.4	20.4	21.2	24.2
Collected sensitive information from online sources	59.2	7.1	22.5	8.2	3.0
Deceived subjects as part of your research	79.9	5.6	10.8	2.6	1.1
Made changes to a study in response to issues raised by an IRB/internal reviewer	48.5	9.8	25.2	12.0	4.5
Inferred characteristics of **groups** based on online data	37.0	4.9	30.2	17.7	10.2

How frequently have you done the following things as part of your study design, data collection, and/or analysis?	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
Shared a dataset with other researchers	31.8%	9.4%	37.8%	11.2%	9.7%
Removed unique individuals from datasets (e.g., only one person from a given country) before sharing data more widely	72.1	2.3	14.5	6.9	4.2
Removed unique individuals from datasets (e.g., only one person from a given country) before analysis	61.5	2.7	23.8	7.7	4.2
Shared raw data with key stakeholders (e.g., community leaders)	73	6	12.4	4.5	4.1
Shared research results with research subjects	27.1	8.8	33.6	20.6	9.9
Notified participants of your data collection and allowed them to opt out	35.1	5.3	12.8	18.9	27.9
Notified participants of your data collection without an opt-out option	85.6	2.3	5.7	3.0	3.4

Notified participants about why you're collecting online data	29.8	3.0	18.5	19.2	29.4
Discussed your data collection with the host site before collecting data	50.0	11.7	21.2	10.2	6.8
Asked colleagues about their research ethics practices	13.6	4.2	42.6	27.2	12.5
Asked your IRB/internal reviewers for advice about research ethics	41.8	8.2	27.6	12.3	10.1
Chosen not to scrape a particular type or category of data because of ethical concerns	50.4	6.4	28.8	8.0	6.4
Argued with your IRB/Internal reviewers about research practices	59.9	5.6	22.1	8.6	3.7
Made code for scraping publicly available (e.g., on Github)	70.8	5.7	13.6	5.3	4.5
Created workarounds on API limitations for data collection	61.4	6.7	16.9	11.2	3.7
I think it is permissible to:	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
Scrape data from online forums	1.1%	4.5%	21.3%	42.2%	31.0%
Reuse data collected for one purpose to answer a new research question	5.2	5.6	18.7	42.7	27.7
Use data from a non-representative online sample	9.8	11.0	22.3	31.8	25.0
Classify individuals or make judgments about participants based on online data	6.4	19.3	19.7	37.9	16.7
Collect sensitive information from online sources	11.4	22.7	29.2	25.8	11.0
Deceived subject as a part of research	26.3	16.2	30.5	22.2	4.9
Use online data for studies where there are minor risks to participants	9.5	14.0	28.4	27.3	20.8
I think researchers should:	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
Remove individuals from datasets upon their request	0.4%	1.1%	6.9%	24.6%	64.5%
Remove unique individuals from datasets (e.g., only person from a given country) before sharing data more widely	4.8	6.3	30.5	27.1	31.2
Remove unique individuals from data sets (e.g., only person from a given country) before analysis	11.6	17.5	38.8	19	13.1
Share raw data with key stakeholders (e.g., community leaders)	13.9	18.8	39.1	20.7	7.5
Notify participants of data collection and allow them to opt out	2.6	7.5	27.6	27.6	34.7
Notify participants of data collection without allowing them to opt out	33.6	34.3	23.4	5.3	3.4
Notify participants about why they're collecting online data	2.3	3.4	28.3	35.1	30.9
Discuss their data collection with the host site before collecting data	5.7	12.3	39.5	28.0	14.6
Ignore a website's Term of Service when necessary to collect data	34.8	31.4	16.7	11.0	6.1
Share research results with research subjects	0.8	1.5	28.7	45.3	23.8
Ask colleagues about their research ethics practices	0.8	0.4	11.7	45.1	42
Ask their IRB/internal reviews for advice about research ethics	1.1	3.4	22.1	38.2	35.1
Argue with their IRB/internal reviewers about research practices	2.7	6.2	41.5	34.6	15.0
Only collect online data when the benefits outweigh the potential harms	5.3	8.3	29.5	33.0	23.9
Think about possible edge cases/outliers when designing studies	0.4	0.4	10.6	43.3	45.2