Big Data and Democracy: Facts and Values

Ashley E. Gorham, *University of Pennsylvania*

ABSTRACT The incongruity between traditional research standards and the realities of "big data" raises novel ethical questions for political science researchers. Data's status as a commodity largely held by private companies only further confuses matters. Drawing on Max Weber, this article attempts to shift the current ethical discussion away from the individualistic concern for privacy toward the more collective framework of democratic values.

Social media data offers an unprecedented opportunity for researchers to access the behavioral traces of individuals. Although such data is an asset for political scientists, the ethics of "big data" research is complicated by the challenges it poses to the traditional ethical concepts of informed consent and harm.[1] Thus far, advocates of regulation have focused on the issue of privacy. However, an individualistic approach is ill-suited to the scale of big data. In pursuit of an alternative regulatory framework for big data research, Max Weber's discussion of facts and values is used to highlight the way in which facts never speak for themselves and, thus, the appropriateness and necessity of locating big data's facts within a value system. This article contends that political scientists should use the democratic value of collective self-determination to guide their approaches to big data's facts.

BIG DATA IN POLITICAL SCIENCE

Excitement over big data has been keenly felt in political science, where social media has facilitated research on a wide range of phenomena. A number of projects using big data have featured collaborations between academics and corporate data scientists, including Bond et al.'s (2012) "A 61-Million-Person Experiment in Social Influence and Political Mobilization" and Bond and Messing's (2015) "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." In the former project, researchers manipulated Facebook users' News Feeds to study political mobilization online (Bond et al. 2012). In the latter project, they used Facebook data on users' 'likes' of "official political pages" to estimate ideology (Bond and Messing 2015, 64). Evans, Cordova, and Sipole (2014) used content analysis to study the use of Twitter by House of Representatives candidates. According to the authors, "67,119 tweets were coded for 1,119 individuals" (Evans, Cordova, and Sipole 2014, 455). They reported distinct "Twitter Style[s]" relative to party affiliation, gender, whether a candidate was an incumbent or not, and the competitiveness of the election (Evans, Cordova, and Sipole 2014). Jamal et al. (2015) used Twitter to examine anti-American Arabic discourse—both in general and following major events. Finding similarly negative attitudes expressed towards Iran, the authors argued that such attitudes reflect a more fundamental dislike of intervention by "societies with different political and social values on the Arab Middle East" (Jamal et al. 2015, 69). Finally, King, Pan, and Roberts (2013, 329) gathered data from 1,382 social media sites in their examination of Internet censorship in China. With the help of "automated text analysis," the authors made the surprising finding that censors tend to allow criticism of the government but remove posts that have collective action potential (ibid., 337, 328).

THE INSUFFICIENCY OF THE CURRENT REGULATORY REGIME

Complicating the big data boom is its novel provenance. Much of big data's *data* is gleaned from user content generated through corporate services. In political science, the majority of data is obtained from social media, mainly Facebook and Twitter. The sheer size of the data puts such research in tension with established research standards, which are based on informed consent and individual calculations of harm. Two research scandals illustrate the insufficiency of these standards in the contemporary information ecosystem. The first is Lewis et al.'s (2008, 331) "Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.com," a study of "the lives and social networks of a complete cohort of college students." The project came under scrutiny for its release of Facebook user data (Zimmer 2010, 316). Zimmer (2010, 319) noted, "[e]ven after removing the names of the subjects, since the dataset still includes race, ethnicity, and geographic data, re-identification of individual subjects remains a real possibility." One researcher defended the project by saying that they had "not accessed any information not otherwise available on Facebook" (Kaufman 2008; quoted in Zimmer 2010, 316, 318, 321). Yet, the publicity of a given profile is relative to the viewer; therefore, the actual level of publicness that individual users consented to remains unclear (Zimmer 2010, 318).

The second example is Kramer, Guillory, and Hancock's (2014) "Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks." For the study researchers from Facebook and Cornell University examined "emotional contagion" by covertly manipulating Facebook users' News Feeds. They found that biasing News Feeds (i.e., making them less positive or

Ashley E. Gorham *is a PhD candidate in the department of political science at the University of Pennsylvania. She can be reached at agorham@sas.upenn.edu.*

© American Political Science Association, 2017
doi:10.1017/S104909651700107X
https://doi.org/10.1017/S104909651700107X Published online by Cambridge University Press

less negative) had an effect on users' emotions as measured by Facebook posts. In an "Editorial Expression of Concern and Correction," the publishing journal's editor-in-chief wrote the following:

> …[q]uestions have been raised about the principles of informed consent and opportunity to opt out in connection with the research in this paper. The authors noted in their paper, "[The work] was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research." When the authors prepared their paper for publication in PNAS [Proceedings of the National Academy of Sciences], they stated that: "Because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell's Human Research Protection Program"…Adherence to the Common Rule is PNAS policy, but as a private company Facebook was under no obligation to conform to the provisions of the Common Rule when it collected the data used by the authors, and the Common Rule does not preclude their use of the data. Based on the information provided by the authors, PNAS editors deemed it appropriate to publish the paper. It is nevertheless a matter of concern that the collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out. (Verma 2014, 10779)

Both the "Tastes, Ties, and Time" and the "Emotional Contagion" studies relied on the second-hand consent derived from Facebook's terms of service.[2] Given the potential for harm that the respective cases involved, the ambiguity of consent is particularly troubling. This is not to say that the researchers intentionally undertook unethical research but rather that, given the scale of big data, it is difficult if not impossible to obtain the traditional IRB benchmarks of informed consent and accurate harm assessment (see boyd and Crawford 2012; Metcalf and Crawford 2016). Social media users occupy a paradoxical position: meticulously self-managing their image in a private type of public sphere that ultimately leaves their information out of their control. They are both intensely self-determining in the production of their data yet largely powerless over the future of the data—that is, how it will be used by the private corporations to whom they submit it and, in this case, how it may be analyzed by academics.

BEYOND THE PRIVACY PARADIGM

Thus far, ethical discussion concerning big data has tended to focus on the issue of privacy. Nissenbaum's (2004, 155) concept of "contextual integrity," understood as "compatibility with presiding norms of information appropriateness and distribution"—or respect for contextual information norms—has been influential in setting the terms of the discussion (see also Barocas and Nissenbaum 2014). Marwick and boyd (2014, 1063) offered a more stringent theory of privacy than Nissenbaum: "networked privacy," which entails "meaningful control over the networked contexts in which the information flows." According to Marwick and boyd, "achieving privacy requires that people have an understanding of and influence in shaping the context in which information is being interpreted." Finally, Crawford and Schultz (2014, 109, 95) advocated "procedural data due process" to address "predictive privacy harms." Such a system would involve notice, a hearing process, impartial adjudication, and judicial review (Crawford and Schultz 2014).

Given the potential for harm that the respective cases involved, the ambiguity of consent is particularly troubling.

Although sympathetic to such efforts, this project pivots on the discussion of privacy toward a broader discussion of values in search of an ethical framework for big data that is more appropriate and more ambitious than the privacy paradigm. Toward this end, the concern with privacy can be seen as gesturing at an unarticulated democratic impulse. In her piece on *Medium*, "What Does the Facebook Experiment Teach Us," danah boyd (2014) wrote: "[m]ore than anything, I want to see users have the ability to meaningfully influence what's being done with their data and I'd love to see a way for their voices to be represented in these processes." In another example, a postdoctoral research fellow at the University of Pennsylvania declined to use data released through a hack of Ashley Madison, a website designed to facilitate affairs. The University's student newspaper, *The Daily Pennsylvanian*, quoted the researcher as saying: "[t]hese people did everything on that website with the assumption that it would be private, and they have in no way said that it's okay to analyze them" (Kannan 2015). The Ashley Madison case is extreme with its overt assumption of privacy by the users; less clear are the semipublic spheres of social media in which the level of privacy assumed may be reduced. That said, the assumption that data will not be taken out of its designated context and analyzed—despite obfuscating user agreements and terms of service—appears reasonable. Whereas the researcher recognized the data as a violation of privacy with the obvious absence of consent, he also was moved by the way in which the individuals lost control over their lives, as they "did everything on that website with the assumption that it would be private." Such a concern is a concern for the democratic.

"Democracy" is generally understood as "the rule of the people," although there is endless debate about what exactly that means. Here, the term is used to refer to the idea of "collective self-determination," a core commitment that underlies—to varying degrees—many modern theories of democracy. Both institutional theories and noninstitutional theories embrace the ideal of collective self-determination. Although it may be most ambitiously pursued in theories of direct democracy it is still present in a diluted form in representative theories. The translation of collective self-determination into practice differs widely across theories, but the centrality of the ideal is constant.

If democracy is understood in terms of collective self-determination, then surreptitious manipulation of user content by researchers represents a conflict with the values of democracy. The "Tastes, Ties, and Time" and "Emotional Contagion" studies are obvious examples, but other less-extreme cases also present problems for democracy. An ethical framework focused on

https://doi.org/10.1017/S104909651700107X Published online by Cambridge University Press

democracy as collective self-determination seeks greater control over data and, in the process, decenters the concepts of informed consent and harm. Democratic values extend beyond the negative freedom of privacy and fold the individualistic notions of informed consent and evaluations of harm into the positive notion of collective self-determination. The democratic ethical framework, therefore, is more appropriate to the current reality in which informed consent and calculations of potential harm are impractical and more ambitious than the privacy paradigm in its granting of power to research subjects.

Although sympathetic to such efforts, this project pivots on the discussion of privacy toward a broader discussion of values in search of an ethical framework for big data that is more appropriate and more ambitious than the privacy paradigm.

WEBER'S FACTS AND VALUES

Max Weber's "Science as a Vocation" provides the foundation for an ethical framework for big data research based on democratic values. The turn to Weber may seem unusual given his distance from the realities of current social science and his ambiguous relationship to the democratic tradition. However, Weber's emphasis on the dependency of scientific research on external systems of value is particularly relevant today as the proliferation of "facts" threatens to obscure the necessity of anchoring value systems. Anderson's (2008) assertion that "[w]ith enough data, the numbers speak for themselves" is only the most famous reminder of the enduring relevance of Weber's insight. Anderson used the example of Venter's "shotgun gene sequencing" to argue for the self-evidence of big data (see Venter 2004). According to Anderson (2008), "high-speed sequencers and supercomputers that statistically analyze the data they produce" allowed Venter to sequence "much of the ocean." He also was able to sequence the "air" and "[i]n the process, he discovered thousands of previously unknown species of bacteria and other life-forms" (ibid.). It is the existence of a "statistical blip" that suggests the existence of "a new species," and correlations "with other sequences" gave Venter information about the nature of the species, as the numbers apparently come to speak for themselves (ibid.).

Although he could not have foreseen big data, Weber's "Science as a Vocation" can be used to challenge Anderson's claim. In this lecture, Weber asserted that while science "presupposes" its worth, it cannot prove it (1989, 18). The relative value of scientific inquiry can be considered only with a specific value system in mind. Weber (ibid.) stated the following:

> …[i]n any scientific work the validity of the rules of logic and method, those general foundations of our orientation in the world, are presupposed. These presuppositions are…the least problematic aspects of science. But science further presupposes that what is produced by scientific work should be *important* in the sense of "being worth knowing." And it is obvious that all of our problems lie here, for this presupposition cannot be proved by scientific means. It can only be *interpreted* with reference to its ultimate meaning, which one must accept or reject according to one's ultimate attitudes towards life.

Thus, numbers—like facts—never speak for themselves. There is always, necessarily, a grounding system of values, or a worldview, that undergirds all of the apparently self-evident truths of big data. It is the scientific worldview that gives value to Venter's discoveries; without it, the meaning of his facts would be uncertain. For Weber, it is in the selection of a research question in particular that the social scientist makes a value judgment, as it entails identifying an object of inquiry as worthy of scientific study. Given the inescapability of value judgments, political scientists would do well to incorporate democratic values into their research. In this way, democratic values could come to regulate big data research in political science.

DEMOCRATIC VALUES AND POLITICAL SCIENCE FACTS

A political science guided by democratic values could take a variety of forms. In *The Filter Bubble*, Pariser (2011) cited the US Department of Housing, Education, and Welfare's "Fair Information Practices" as a model for contemporary regulation of information. The document calls for transparency of collection and use, the limitation of information to its intended use, a correction or amendment process, and standards for security (Pariser 2011, 238–9). Pariser also proposed the development of "an agency to oversee the use of personal information" (ibid., 241). Research based on Pariser's model would increase transparency regarding methods and goals, restrict data to the expressed purpose for which it was collected, create a method for correcting erroneous data, and establish standards for data security. Researchers also could choose to limit their research to subjects that have expressly opted in to their projects, rather than relying on the implied consent gained from terms-of-service agreements. Finally, "dynamic consent" and "expiry dates for consent" could help counteract the despotism of implied consent (Whitley, Kanellopoulou, and Kaye 2012; Custers 2016).

If democracy is understood in terms of collective self-determination, then surreptitious manipulation of user content by researchers represents a conflict with the values of democracy.

Although these projects embody democratic values, a truly democratic research program would go beyond restrictions guided by the democratic ideal of self-determination to embrace positive, *collective* self-determination in practice. Toward this end, researchers could work with social media companies to offer

https://doi.org/10.1017/S104909651700107X Published online by Cambridge University Press

surveys, referendums, and even town halls to discuss proposed research projects. By lot or through online elections, representatives could be chosen to debate and offer their input on specific projects. Collectively, social media communities could decide whether they approved of individual research projects, whether they would like to participate in projects, and whether they approved of political scientists conducting research on social media in general. Whatever form these practices took, they all would afford subjects greater control over how their social media data would be used by researchers.

The call for political science to be mindful of democratic values in its work with big data highlights the need for a larger discussion that goes beyond the ethics of academic research to the question of the democratization of data in general. This necessarily would involve exploring what data *is* and how it relates to other commodities. As increasingly more data is produced on social media, the digital shadows of individuals' lives—or "shadow selves"—expand. How much control individuals will have over those shadow selves will be an important question and will depend largely on how data is conceptualized. Such a project is beyond the scope of this article, but it is hoped that this may be a step in that direction. ■

NOTES

1. The most common definition of big data is "high volume, high velocity, and/or high variety information assets" (Gartner 2017). However, as Kitchin and McArdle (2016, 9) noted, "Big Data do no all share the same characteristics and...there are multiple forms of Big Data." Eschewing technical terms, Clark and Golder (2015, 66) seem to express the general understanding of the term within political science with their conceptualization of big data as "the idea that technological innovations such as machine learning have allowed scholars to gather either new types of data, such as social media data, or vast quantities of traditional data with less expense."
2. In the case of Lewis et al. (2008), the university also explicitly agreed to the study and cooperated with the researchers.

REFERENCES

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired*, June 23. Available at www.wired.com/2008/06/pb-theory.

Barocas, Solon, and Helen Nissenbaum. 2014. "Big Data's End Run around Anonymity and Consent." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, ed. Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 44–75. New York: Cambridge University Press.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489 (7415): 295–8. DOI: 10.1038/nature11421.

Bond, Robert, and Solomon Messing. 2015. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109 (1): 62–78. DOI: 10.1017/S0003055414000525.

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15 (5): 662–679. DOI: 10.1080/1369118X.2012.678878.

boyd, danah. 2014. "What Does the Facebook Experiment Teach Us? Growing Anxiety About Data Manipulation." *Medium*, July 1. Available at https://medium.com/message/what-does-the-facebook-experiment-teach-us-c858c08e287f.

Clark, William Roberts, and Matt Golder. 2015. "Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?: Introduction." *PS: Political Science & Politics* 48 (1): 65–70. DOI: 10.1017/S1049096514001759.

Crawford, Kate, and Jason Schultz. 2014. "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms." *Boston College Law Review* 55 (1): 93–128. Available at http://lawdigitalcommons.bc.edu/bclr/vol55/iss1/4.

Custers, Bart. 2016. "Click Here to Consent Forever: Expiry Dates for Informed Consent." *Big Data & Society* 3 (1): 1–6. DOI: 10.1177/2053951715624935.

Evans, Heather K., Victoria Cordova, and Savannah Sipole. 2014. "Twitter Style: An Analysis of How House Candidates Used Twitter in Their 2012 Campaigns." *PS: Political Science & Politics* 47 (2): 454–62. DOI: 10.1017/S1049096514000389.

Gartner. 2017. "IT Glossary: Big Data." Available at www.gartner.com/it-glossary/big-data.

Jamal, Amaney A., Robert O. Keohane, David Romney, and Dustin Tingley. 2015. "Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses." *Perspectives on Politics* 13 (1): 55–73. DOI: 10.1017/S1537592714003132.

Kannan, Vibha. 2015. "Penn Researcher Refuses to Analyze Ashley Madison Data." *Daily Pennsylvanian*, August 30. Available at www.thedp.com/article/2015/08/penn-researcher-refuses-to-analyze-ashley-madison-data.

Kaufman, Jason. 2008. "Michael—We Did Not Consult..." Michaelzimmer.org. October 2, 9:25 p.m. Blog comment. Available at www.michaelzimmer.org/2008/09/30/on-the-anonymity-of-the-facebook-dataset.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107 (2): 326–43. DOI: 10.1017/S0003055413000014.

Kitchin, Rob, and Gavin McArdle. 2016. "What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets." *Big Data & Society* 3 (1): 1–10. DOI: 10.1177/2053951716631130.

Kramer, Adam D. I., 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): 8788–90. DOI: 10.1073/pnas.1320040111.

Lewis, Kevin, 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): 330–42. DOI: 10.1016/j.socnet.2008.07.002.

Marwick, Alice E., and danah boyd. 2014. "Networked Privacy: How Teenagers Negotiate Context in Social Media." *New Media & Society* 16 (7): 1051–67. DOI: 10.1177/1461444814543995.

Metcalf, Jacob, and Kate Crawford. 2016. "Where are Human Subjects in Big Data Research? The Emerging Ethics Divide." *Big Data & Society* 3 (1): 1–14. DOI: 10.1177/2053951716650211.

Nissenbaum, Helen. 2004. "Privacy as Contextual Integrity." *Washington Law Review* 79: 119–58.

Pariser, Eli. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. New York: Penguin.

Venter, J. Craig, et al. 2004. "Environmental Genome Shotgun Sequencing of the Sargasso Sea." *Science* 304 (5667): 66–74. DOI: 10.1126/science.1093857.

Verma, Inder M. 2014. "Editorial Expression of Concern and Correction: Experimental Evidence of Massivescale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences* 111 (29): 10779. DOI: 10.1073/pnas.1412469111.

Weber, Max. 1989. "Science as a Vocation." In *Max Weber's "Science as a Vocation,"* ed. Peter Lassman and Irving Velody with Herminio Martins; translated by Michael John, 3–31. Boston: Unwin Hyman.

Whitley, Edgar A., Nadja Kanellopoulou, and Jane Kaye. 2012. "Consent and Research Governance in Biobanks: Evidence from Focus Groups with Medical Researchers." *Public Health Genomics* 15 (5): 232–42. DOI: 10.1159/000336544.

Zimmer, Michael. 2010. "'But the Data Is Already Public': On the Ethics of Research in Facebook." *Ethics and Information Technology* 12 (4): 313–25. DOI: 10.1007/s10676-010-9227-5.

https://doi.org/10.1017/S104909651700107X Published online by Cambridge University Press

2017 Ralph Bunche Summer Institute



2017 APSA RBSI Scholars, Faculty, & Staff

Named in honor of the 1950 Nobel Peace Prize winner, former APSA President, and the first African American to receive a PhD in political science, Dr. Ralph J. Bunche, the RBSI encourages students to pursue academic careers in political science. During the program, RBSI Scholars take graduate level political science and methods courses and engage in original research, culminating in a research paper and poster presentation at the APSA Annual Meeting. To date, there are over 500 RBSI Alumni, many of whom are on faculty at colleges and universities across the country. Now in its 31st year, the RBSI program is held at Duke University, under the direction of Dr. Paula D. McClain. To learn more about the 2017 APSA RBSI Scholars, visit us online at www.apsanet.org/rbsi.

Congratulations to the APSA Ralph Bunche Summer Institute!

Photo Credit: Shaun King, *Duke University, Department of Political Science Multimedia Specialist*

https://doi.org/10.1017/S104909651700107X Published online by Cambridge University Press