COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 231–238
DOI: 10.17645/mac.v8i3.3163

Article

Open-Source's Inspirations for Computational Social Science: Lessons from a Failed Analysis

Nathaniel Poor

Underwood Institute, Cambridge, MA 02139, USA; E-Mail: natpoor@gmail.com

Submitted: 15 April 2020 | Accepted: 7 July 2020 | Published: 13 August 2020

Abstract
The questions we can ask currently, building on decades of research, call for advanced methods and understanding. We now have large, complex data sets that require more than complex statistical analysis to yield human answers. Yet as some researchers have pointed out, we also have challenges, especially in computational social science. In a recent project I faced several such challenges and eventually realized that the relevant issues were familiar to users of free and open-source software. I needed a team with diverse skills and knowledge to tackle methods, theories, and topics. We needed to iterate over the entire project: from the initial theories to the data to the methods to the results. We had to understand how to work when some data was freely available but other data that might benefit the research was not. More broadly, computational social scientists may need creative solutions to slippery problems, such as restrictions imposed by terms of service for sites from which we wish to gather data. Are these terms legal, are they enforced, or do our institutional review boards care? Lastly—perhaps most importantly and dauntingly—we may need to challenge laws relating to digital data and access, although so far this conflict has been rare. Can we succeed as open-source advocates have?

Keywords
computational social science; fandom; games; online community; open source; Reddit

Issue
This article is part of the issue "Computational Approaches to Media Entertainment Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences, Germany), Tim Wulf (LMU Munich, Germany) and M. Rohangis Mohseni (TU Ilmenau, Germany).

1. Introduction

This article uses an autoethnographic approach to explore issues I encountered on a project about a gaming community and its online embodiment. I discuss how lessons from the open-source community could have helped me and, more broadly, can help computational social science (CSS) move past current tensions with application programming interface (API) availability, commercialism, and academic research for the greater social good. My study, part of a larger work, used data on more than 2.25 million posts to the online forum site Reddit, captured over two one-year periods, accessed via an API offered through Pushshift. The hypothesis for that project was that users move from one community to another when a new game in a series comes out (hence the two one-year time periods, related to two game releases and opportunities for community movement). My initial findings, despite a large amount of data and relevant theory, was that users do not in fact move from community to community—a null finding.

This current article explores how this null finding could have happened, even with so much data and good theory, and proposes a way forward for CSS as a whole in light of such occasional, but potentially enlightening, problems.

First, I touch on both open-source software and CSS as ideological in nature, in order to frame the autoethnographic case study of my initially failed research project. Then I discuss my study by explaining the site (Reddit) and the API (Pushshift). I present the topic in terms of the specific game series (Bethesda's Elder Scrolls franchise)

and the theories involving online communities and fans. I then move on to the analysis of why the study did not work and the solutions I arrived at.

These solutions serve as a jumping-off point for the main idea of this article. The solutions for my specific research problems can be contextualized within challenges that researchers have identified for CSS and that must be overcome (e.g., Bruns, 2019; Freelon, 2018; Halavais, 2019). But we can draw lessons from the history and current mature state of the open-source software ecosystem to help us move forward. The open-source world is an environment with a wide mix of commercial and freely available items, third parties who add value to information, teams of people with different expertise, and legal hurdles that have had to be overcome, similar to the CSS world itself.

1.1. Open-Source Software

Open-source software (DiBona, Ockman, & Stone, 1999; Raymond, 1999) is software for which the code (and the software itself) is available for anyone to copy (*libre*) as well as available for free (*gratis*). Sometimes it is referred to as 'free and open-source software,' or FOSS, although opinions differ about the definitions important to those involved in the effort (DiBona et al., 1999). Beyond this, the idea is that the code can be freely modified so that users can fix bugs or customize it as they need. The source code is available; it is open-source. This openness contrasts with commercial software, for which users must pay and for which they cannot access the source code. Unlicensed copies and the hacking of commercial software to get at its code are generally illegal. Note that all of these considerations are ideological, economic, and political. These are not just technologies, but rather they are sociotechnical systems.

Open-source is by now well established and does not garner the attention it once did. Many academics are aware of open-access journals and the open-access licenses governing articles in such journals. These stem from the licensing ideas in open-source software, which utilize copyright law to allow copying with certain requirements (such as attribution) instead of disallowing it. Open-access journals are in some ways the journal equivalent of open-source software. Open-source advocates have carefully reconsidered copyright within the existing legal framework to essentially turn the concept on its head: instead of copyright, there is copyleft. That is, instead of using copyright laws to restrict copying of works, advocates realized that licensing laws could be used to restrict the restriction that would put those works under lock and key.

Open-source, besides a type of legal license or ideological stance, is also a way of working. Many people freely donate time and expertise to work on large, distributed open-source projects. But not all licenses require that the end product always be free. Some allow modifications that users or open-source-related companies can charge for. Other companies, such as O'Reilly, publish guidebooks to open-source software, and they charge for those books just as they would charge for any other book.

People undertaking open-source work often use open-source tools to make more open-source code in turn, such as using open-source text editors and compilers to make open-source computer programs. Some CSS tools, such as Python, R, and some database applications, are open-source.

The software world is a mixed environment, much like the CSS world of accessing data and research. Some source code (or data) is free, other source code is not, and one must pay for access or not access it at all. Some people add value to the data, some people work on data for free, and teams of people with a variety of relevant skills often work on projects.

1.2. CSS as Ideology

CSS, generally, is the large-scale analysis of digital data relating to human behavior (Lazer et al., 2009). The size of the data set needed to achieve the CSS label varies depending on the perspective of the viewer; attainable sizes (in terms of collection, storage, and analysis) have increased over time. Data complexity also may present a challenge to deciding what is and what is not CSS, as may the type of analysis used.

The advent of CSS can, in hindsight, be seen as a sensible sociotechnical response to improvements in several technological areas: local computing power, easier programming languages, greater internet speeds, data access via APIs, and a greater number of people interacting with a greater amount of digital material online. As part of the social sciences, CSS relies on the belief that things can and should be measured, and that those measurements can accurately measure what we are trying to measure (Bulmer, 2001). This is ideological. Some of the technological advances needed for CSS are ideological as well: the arguments that computing languages *should* be easier than, for example, C, or that data *should* be collected and then made available by API. This point should not be overlooked, because sites from which we may want data can have restrictive, vague, and problematic terms of service that perhaps researchers should ignore in light of a greater social good (Fiesler, Beard, & Keegan, 2020). Additionally, when APIs are shut down and data access is curtailed, some CSS studies become impossible or at least difficult to undertake (Bruns, 2019; Freelon, 2018).

When the API is fronting data that has commercial value, as with Facebook and Twitter, or if the data come with privacy concerns, making the data available freely and for free becomes problematic in different ways (Bruns, 2019; Fiesler et al., 2020; Halavais, 2019). Much like with the open-source community in its early days of growing popularity, a tension exists between those who want information to be freely available for a greater

social good and those who want to commercialize it (DiBona et al., 1999; Raymond, 1999). This tension can be approached by understanding commercial restrictions and economic gain for the few, on the one hand, and academic access and research for the greater good, on the other—although both sides share data privacy concerns, albeit for different reasons, working toward different outcomes.

The API, however, was not the problem I encountered in my research's null finding. The data were available, in fact, because one person believed they should be available. Without that API, my project might have been impossible.

2. The Research Project

2.1. Reddit and Pushshift

Reddit is a website that serves as an online space for thousands of different forums, called subreddits, many of which function as online communities (Panek, Hollenbach, Yang, & Rhodes, 2018) and, importantly, as online fan communities (Gunderman, 2020). Some subreddits receive hundreds or thousands of posts per day. Founded in 2005, Reddit is similar to, and draws from, older online bulletin-board systems like Usenet, AOL, Slashdot, and modem-based BBS systems. The people who run Reddit generally take a hands-off approach to site governance, which has led to some problems (Massanari, 2017). But the users make each specific subreddit and determine its rules. Like many online spaces, users may create a username (and thus an identity of sorts) if they wish and may post to whichever subreddits they like, or they might just lurk and read posts without commenting. But if they do post, they create both text (the post) and associated metadata (such as who posted, where, and when). These digital trace data were what I wanted, but Reddit does not make it easily available. Instead, an individual, Jason Baumgartner, has taken it upon himself to collect all of it, billions of posts, and make it available via an API at the website Pushshift.io (Gaffney & Matias, 2018). Note that he does this for free and solicits donations to help the effort.

2.2. The Elder Scrolls, Fans, and Online Gaming Communities

The Elder Scrolls franchise is a series of fantasy adventure computer games reminiscent of *The Hobbit* (Tolkien, 1937), with elves and wizards and magic swords. The initial game in the series, The Elder Scrolls: Arena (Bethesda Softworks, 1994), was released in 1994. The Elder Scrolls V: Skyrim (Bethesda Softworks, 2011) was released in November 2011, was remastered in 2016 for newer game consoles, and was recoded and released for the Nintendo Switch in 2017. In short, Skyrim is extremely popular and has sold millions of copies worldwide. A massively multiplayer game set in the Elder Scrolls universe, The Elder Scrolls Online (Bethesda Softworks, 2014), similar to the better-known World of Warcraft (Blizzard Entertainment, 2004), was released in 2014. The releases of these two games—Skyrim and The Elder Scrolls Online—were the points in time in which I was interested.

Many people who buy Elder Scrolls games are more than just purchasers or players of the game. They are fans, as is true of many people and many cultural products (Fiesler, 2007; Jenkins, 1992). Fans, and people more generally, form communities and online communities; currently many communities have both online and offline components to varying degrees (Poor & Skoric, 2014; Wellman, Boase, & Chen, 2002). Fan communities can be robust and can survive moves from one platform to another (Fiesler, Morrison, & Bruckman, 2016; Pearce, 2009).

Some fans of games like the Elder Scrolls go beyond just buying the game. They participate in actively creating and changing the game worlds, such as by coding modifications, or mods, to those games when possible (usually on the Windows/PC platform). Game modders form their own subculture within gaming fans of a game; and the many necessary interactions among modders can lead to strong community ties (Poor, 2014). More generally, fans are also well known for creating fictional stories about the objects of their fandom, called fanfic (Jenkins, 1992).

It is through communication—for fans, perhaps communication takes the form of fanfic, discussion of how to make a mod, or discussion of the game in general—that humans form community (Carey, 1989; Dewey, 1927; Iyer, Cheng, Brown, & Wang, 2020). This phenomenon is not found just in our online behavior (Kraut & Resnick, 2011). It is a fundamental capability that evolved in us over millions of years (Gamble, Gowlett, & Dunbar, 2014; Tomasello, 2010). Fandom is not solely denoted by communication: Actually buying the objects related to that fandom is an important and a heavily intertwined part of how fan identities are established and maintained (Hills, 2003). To some extent, fans are "ideal consumers [who] automatically buy the latest works" (Cavicchi, 1998, p. 62).

Altogether, being a fan of one Elder Scrolls game or of the franchise overall, buying the new game in the franchise upon its release, and then discussing it with other fans on Reddit seems like a sensible path to many fans of the series based on the above-mentioned theory. That was my main hypothesis in my failed study.

2.3. Data and Results

Using Python, I scraped the third-party Reddit API at Pushshift, ending up with data for more than 2.25 million posts spanning two full years. For the first one-year period when Skyrim (Bethesda Softworks, 2011) was released, I obtained data on 979,582 posts: who posted, when, and to which of the several game-related sub-

reddits. To study correlations in posting behavior, I winnowed it down to the three months before and the three months after the new game was released, using data on 772,873 posts. For the second new game (The Elder Scrolls Online [Bethesda Softworks, 2014] superseded Skyrim as the newest game in the franchise), I scraped data on 1,296,146 posts, which I similarly winnowed down to data on 861,040 posts spanning the three months before and the three months after its release.

I ran correlations between number of posts on one subreddit during the three months before a game release and number of posts on another subreddit during the three months after the same game release, per user. This captured before-release and after-release posting behavior. The first correlation, from The Elder Scrolls IV: Oblivion (Bethesda Softworks, 2006) to The Elder Scrolls V: Skyrim (Bethesda Softworks, 2011), was 0.16. The second, from Skyrim to The Elder Scrolls Online (Bethesda Softworks, 2014), was 0.04. Clearly, this null result was headed for the file drawer (Rosenthal, 1979). How do you get a null result with 2.25 million data objects and good theory?

2.4. Problems and Solutions

Perhaps I did not have quite the right data. Maybe I needed more specific sales data or player data, which exist but cannot be accessed. Maybe I should have included the text of the posts, not just the metadata, and performed textual analysis to gauge the sentiment of users toward the new game, or else to determine where their level of devotion lay. Were they fans of the old game specifically, or of the series generally? Possibly I had aggregated data at an analytically unsound level, encountering a Simpson's paradox as can befall such work on Reddit (Lerman, 2018).

Perhaps I should have added surveys to the digital data, as some researchers have suggested (Stier, Breuer, Siegers, & Thorson, 2019). Perhaps I was too narrow in my focus, looking only at the fans or users as they participated on Reddit but missing their activity on other platforms (Menchen-Trevino, 2013). Researchers have noted how Skyrim (Bethesda Softworks, 2011) fans (Puente & Tosca, 2013) and people in general (Baym, 2007) use more than one online space for their online activity.

Possibly my analysis was not very good. Initially I had wanted a path-analysis model (given how people use multiple subreddits over time); but after consultation with a colleague I went with much simpler correlations. Maybe I had the wrong theory—perhaps consumerism was not the driving force here (indeed, community attachment turned out to be the story). How could I have avoided this null finding?

My eventual solution was to add a co-author who has more expertise in the platform and qualitative methods than I do, and who also works in game studies. Her partner and occasional co-author is also a data scientist, one far more skilled than I am at using big data techniques. We ended up looking at the data again and reinterpreting it, which led us to consider new theories and, in turn, led us to go over the data again in an iterative approach. Our further thinking led us to consider how to get data from multiple platforms about the players of interest, which in some cases seems impossible because those data aren't made available by the companies in question. Each of these solutions, which worked for this specific project, should also be viewed in a wider and more general context within academic research while keeping in mind the issues faced by open-source advocates and practitioners.

3. Solutions in a Larger Context

In summary, the solutions to the challenges we had to resolve were as follows: 1) Use an interdisciplinary team with expertise across methods, theory, and topic, 2) use a continually iterative understanding of the theory, data, methods, and findings, and 3) acknowledge the limitations that may stem from a data environment that includes free data and protected or unattainable data.

More broadly, two additional issues hover over research, issues with which both computational social scientists and open-source advocates must deal. They must: 1) Work within a potentially restrictive legal environment in a creative manner to move work forward (for open-source there is copyleft, for CSS there is working around or through terms of service [Fiesler et al., 2020; Halavais, 2019]), and 2) challenge that restrictive legal environment directly (as suggested by Puschmann [2019], and as successfully undertaken by Christian Sandvig at the University of Michigan [American Civil Liberties Union, 2020]).

3.1. Team Work

The advanced questions we can ask now, building on decades of research and methods, call for advanced methods and understanding. Advanced methods require more than just large amounts of raw computing power; they call for both quantitative and qualitative methodological approaches (Menchen-Trevino, 2013), a broad understanding of theories, topical expertise, and computational skills. In short, CSS done well requires teams (Lazer et al., 2009).

Large, complex data sets require more than complex statistical analysis to come to human answers. Humans are messy and beautiful, and qualitative methods are much better positioned to capture and understand the beautiful mess than are quantitative methods (Law, 2004). Used together, however, they can present enlightening pictures of human behavior—hence the somewhat recent move toward mixed methods (Creswell, 2009) and the understanding that both types of methods complement each other (Strauss & Corbin, 1998).

This necessary variety is similar to work in FOSS, where large projects require teams with varied exper-

tise depending on the project, which can include coding languages (Python), analysis languages (R), databases (MongoDB, MySQL), all parts of an operating system (Linux), web browsers (Firefox), and graphics software (GIMP).

3.2. Continually Iterative

Open-source projects take time to put together, test, and release. Creators might roll out new releases. That is, people have constantly worked on them, considered issues, and revised along the way. Although an end result emerges, the ongoing process is a vital part of the overall effort. The same is true, or should be, for many CSS projects.

The way we write up research makes it seem as if it has followed a nice, linear narrative that happens to fit the journal article format rather well. First, we read some literature, and that makes us think of some hypotheses and some methods to test them, and only then do we happen to find (or create) the perfect data and arrive at publishable results. But much research does not work this way. Preregistration is one important step for certain types of studies (Nosek, Ebersole, DeHaven, & Mellor, 2018); it is an important acknowledgment of this issue and of the file drawer problem (Rosenthal, 1979).

Researchers working on qualitative and combined methodologies have engaged with this issue and have espoused the usefulness of an iterative approach (Davidson, Edwards, Jamieson, & Weller, 2019; Muller, Guha, Baumer, Mimno, & Shami, 2016). For instance, in a grounded theory approach, researchers might iteratively build categories from data, revisiting the data again and again as they recognize more categories (Strauss & Corbin, 1998). This process can be especially important for CSS projects, where understanding the often large and diverse data set takes more than one pass.

3.3. Mixed Data Environment

In addition to mixed-methods approaches with quantitative and qualitative data (Creswell, 2009; Nelson, 2017), computational social scientists work in an environment with a variety of available data. Some are free, such as Wikipedia data. Some data have slightly restrictive license requirements, such as some APIs that require registration and user tokens or authentication (Bruns, 2019). Some data require payment before one can access them. Some data repositories, such as Pushshift, hope for donations. Other data might be curated (Gruzd, 2016). Some data have been hacked and released (Poor, 2017) or unwisely released (Resnick, 2016). Other data are only available to in-house researchers, or not at all.

This situation is similar in part to the environment in which FOSS programmers work, in that code is available under a variety of licenses, ranging from free to restricted to simply unavailable.

3.4. Creative Solutions

Turning copyright on its head into copyleft was a highly creative solution to the problems that FOSS advocates wanted to address. Copyright, at least in the US context—even with fair use exceptions—is almost always used to deny people the right to copy. FOSS advocates needed a way to allow copying, but with certain restrictions, allowances, and requirements, such as making the changed code available for free or giving credit to previous coders.

Researchers in CSS need access to data but cannot always get it (Bruns, 2019; Freelon, 2018). One problem can arise from a site's terms of service, which may disallow data scraping even when it is technologically possible (Fiesler et al., 2020; Halavais, 2019). Although researchers can ignore the terms of service and scrape a website anyway—hoping to avoid rate limits, throttling, and getting blocked completely—researchers can also approach users directly (Halavais, 2019). Another approach could be to claim fair use doctrine (under American law) as conferring a right to copy the information for academic research purposes.

3.5. Challenge the Existing Structure

FOSS advocates took on the commercial software industry and succeeded. FOSS software is widespread, from some of our own CSS tools to the Linux operating system to the Apache web server (which has been the most widely used web server for many years). Even some large for-profit corporations like IBM support FOSS. But the effort has not always been easy, and the FOSS world is no stranger to lawsuits in which FOSS licensing terms have been upheld by the courts or where settlements have been reached in favor of the FOSS litigant (e.g., Neuburger, 2009; Smith, 2009; Stricklett, 2020).

Most professors, however, are not used to acting as social agitators or legal advocates, except perhaps in social work or law. But FOSS advocates needed to face this challenge in order to legitimize their work. Similarly, academic researchers may need to face established legal and economic structures in order to legitimize important research efforts, although to date few have (e.g., American Civil Liberties Union, 2020). One approach is to educate and work with legislative bodies (Puschmann, 2019); another is to use the courts. Although professors are supposed to have some legal protections for their work, few may want to risk a legal challenge with an unclear outcome to pursue one research project when they could do other research instead. This reluctance may have to change. Additionally, institutions such as universities and associations (e.g., the Association for Computing Machinery) have greater resources compared to individuals and may have to lead in this area, whether pushed by members or guided by leadership.

Note that such a legal effort, like FOSS's work and its now-established legal precedent, requires a team of

people. Taking a legal effort to court is daunting. The recent and successful challenge to the Computer Fraud and Abuse Act in the US was led by Christian Sandvig (American Civil Liberties Union, 2020), but Sandvig is not a lawyer. He is a communication studies professor at the University of Michigan. Other plaintiffs in that legal action included researchers from the University of Illinois and Northeastern University, as well as one commercial publisher, although the American Civil Liberties Union and its legal team were behind the case, and it was Sandvig who gave his name to the case: *Sandvig v. Sessions* (2018). The findings of the case are narrow, but they may eventually be seen as an important and inspirational first step on a longer journey.

3.6. Remaining Issues

One place where this comparison of FOSS with CSS falls short is that in the open-source world, the programmers make the code, the content, and the data they work with. In CSS, researchers mostly do not make the data; they scrape it or use an API to access it. Hence the idea of researchers creating data, via surveys and interviews perhaps in a mixed-methods study, may be of great importance (Bruns, 2019; Freelon, 2018). Beyond creating data in this way, the next step parallel to the FOSS example is to make the data available. Indeed, the idea of making data available has been gaining traction among researchers, although important ethical issues arise when sharing human-centric digital data (Fiesler et al., 2020), which is a problem when considering whether social media and other companies should make data available at all.

Could academics and industry work together (Puschmann, 2019)? Some commercial software companies do support FOSS work by their employees, but the academy/industry relation in social media is somewhat different. Some companies such as Facebook have their own internal research departments, and overall, their motivation to share data with academics is unclear. Perhaps their data drive their advertising revenue and are locked down, but using employee time and company resources to aid academic research also subtracts from the bottom line.

4. Conclusions and the Way Forward

The initial findings for my study—that Reddit users were not moving from the subreddit for the old Elder Scrolls game to the subreddit for the new Elder Scrolls game—was accurate if unexpected, despite having data on more than 2,25 million posts for the overall study across two one-year periods. The literature and my own personal experience had led to a well-grounded, albeit unsupported, hypothesis. My mistake was that I had tried to do too much on my own. In the FOSS world, there is a saying: 'with enough eyeballs, all bugs are shallow.' I needed a team of experts. Eventually we iterated over the research from start to finish, and then over it again (theory, data, methods, conclusions), reaching better conclusions. We debated what we could do with the data we could access, and how we could access it, and wondered about the data we could not get. This picture is familiar to FOSS practitioners.

CSS practitioners may not be prepared to be part of a movement or a revolution in the way that many FOSS advocates have seen themselves (DiBona et al., 1999), but some are already taking steps in that direction. We have to engage with potential legal issues with the terms of service, and we might choose to ignore the terms of service altogether. We might have to deal with intransigent institutional review boards which are there not to help us but instead to protect the university (Halavais, 2019), and whose word is essentially university law. We may have to work with legislative governmental bodies, as suggested by Puschmann (2019). Attempting to challenge national laws through the courts and reframe them in our favor is a daunting step that requires several years, a solid case, and a top-notch legal team (American Civil Liberties Union, 2020). So far, few have dared go this route. More may have to do so if we are to flourish as a field. The alternative is to sit passively while big data companies profit from their advertising and guide the laws in ways that favor their own commercial interests, to the disadvantage of academic research for the greater social good.

Acknowledgments

I thank Dr. Johannes Breuer, Dr. Tim Wulf, and Dr. Rohangis Mohseni for their efforts in organizing this thematic issue and for their encouragement for this submission, and the anonymous reviewers who helped guide this article toward its current form.

Conflict of Interests

The author declares no conflict of interests.

References

American Civil Liberties Union. (2020). Federal court rules 'big data' discrimination studies do not violate Federal anti-hacking law. *American Civil Liberties Union*. Retrieved from https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti

Baym, N. (2007). The new shape of online community: The example of Swedish independent music fandom. *First Monday*, *12*(8).

Bethesda Softworks. The Elder Scrolls IV: Oblivion [Video game]. (2006). Rockville, MD: Bethesda Softworks.

Bethesda Softworks. The Elder Scrolls Online [Video game]. (2014). Rockville, MD: Bethesda Softworks.

Bethesda Softworks. The Elder Scrolls V: Skyrim [Video game]. (2011). Rockville, MD: Bethesda Softworks.

Bethesda Softworks. The Elder Scrolls: Arena [Video game]. (1994). Rockville, MD: Bethesda Softworks.

Blizzard Entertainment. World of Warcraft [Video game]. (2004). Irvine, CA: Blizzard Entertainment.

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information Communication and Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Bulmer, M. (2001). Social measurement: What stands in its way? *Social Research*, *68*(2), 454–480.

Carey, J. W. (1989). *Communication as culture: Essays on media and society*. New York, NY: Routledge.

Cavicchi, D. (1998). *Tramps like us: Music and meaning among Springsteen fans*. Oxford: Oxford University Press.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. London: Sage.

Davidson, E., Edwards, R., Jamieson, L., & Weller, S. (2019). Big data, qualitative style: A breadth-and-depth method for working with large amounts of secondary qualitative data. *Quality and Quantity*, *53*(1), 363–376. https://doi.org/10.1007/s11135-018-0757-y

Dewey, J. (1927). *The public and its problems*. Denver, CO: Swallow Press.

DiBona, C., Ockman, S., & Stone, M. (Eds.). (1999). *Open-sources: Voices from the open-source revolution*. Sebastopol, CA: O'Reilly.

Fiesler, C. (2007). Everything I need to know I learned from Fandom: How existing social norms can help shape the next generation of user-generated content. *Vanderbilt Journal of Entertainment and Technology Law*, *10*, 729–762.

Fiesler, C., Beard, N., & Keegan, B. (2020). No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of the 2020 International Conference on Web and Social Media*.

Fiesler, C., Morrison, S., & Bruckman, A. S. (2016). An archive of their own: A case study of feminist HCI and values in design. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2574–2585). New York, NY: ACM Press. https://doi.org/10.1145/2858036.2858409

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1007/978-981-13-0402-6_6

Gaffney, D., & Matias, J. N. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLOS ONE*, *13*(7). https://doi.org/10.1371/journal.pone.0200162

Gamble, C., Gowlett, J., & Dunbar, R. (2014). *Thinking big: How the evolution of social life shaped the human mind*. London: Thames & Hudson.

Gruzd, A. (2016). Social media data stewardship. *Social Media Lab*. Retrieved from https://socialmedialab.ca/2016/03/21/defining-social-media-data-stewardship-smds

Gunderman, H. C. (2020). View of fan geographies and engagement between geopolitics of Brexit, Donald Trump, and Doctor Who on social media. *Transformative Works and Cultures*, *32*. https://doi.org/10.3983/twc.2020.1675

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information Communication and Society*, *22*(11), 1567–1581. https://doi.org/10.1080/1369118X.2019.1627386

Hills, M. (2003). *Fan cultures*. New York, NY: Taylor & Francis.

Iyer, S., Cheng, J., Brown, N., & Wang, X. (2020). When does trust in online social groups grow? In *Proceedings of the fourteenth international AAAI conference on web and social media* (pp. 283–293). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Jenkins, H. (1992). *Textual poachers: Television fans and participatory culture*. New York, NY: Routledge.

Kraut, R. E., & Resnick, P. (2011). *Building successful online communities*. Cambridge, MA: MIT Press.

Law, J. (2004). *After method: Mess in social science research*. New York, NY: Routledge.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L. L., Brewer, D., . . . Alstyne, M. V. (2009). Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lerman, K. (2018). Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science*, *1*(1), 49–58. https://doi.org/10.1007/s42001-017-0007-4

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346.

Menchen-Trevino, E. (2013). Collecting vertical trace data: Big possibilities and big challenges for multi-method research. *Policy & Internet*, *5*(3), 328–339. https://doi.org/10.1002/1944-2866.POI336

Muller, M., Guha, S., Baumer, E. P. S., Mimno, D., & Shami, N. S. (2016). Machine learning and grounded theory method: Convergence, divergence, and combination. In *Proceedings of the 19th international conference on supporting group work* (pp. 3–8). https://doi.org/10.1145/2957276.2957280

Nelson, L. K. (2017). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Neuburger, J. (2009). Jacobsen v. Katzer: Open-source Software project gains key rulings in copyright infringement litigation. *New Media and Technology Law Blog*. Retrieved from https://newmedialaw.proskauer.com/2009/12/16/jacobsen-v-katzer-open-source-software-project-gains-key-rulings-in-copyright-infringement-litigation

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606.

Panek, E., Hollenbach, C., Yang, J., & Rhodes, T. (2018). The effects of group size and time on the formation of online communities: Evidence from Reddit. *Social Media + Society*, *4*(4), 1–13. https://doi.org/10.1177/2056305118815908

Pearce, C. (2009). *Communities of play*. Cambridge, MA: MIT Press.

Poor, N. (2014). Computer game modders' motivations and sense of community: A mixed-methods approach. *New Media and Society*, *16*(8), 1249–1267. https://doi.org/10.1177/1461444813504266

Poor, N. (2017). The ethics of using hacked data: Patreon's data hack and academic data standards. In M. Zimmer & K. Kinder-Kurlanda (Eds.), *Internet research ethics for the social age: New challenges, cases, and contexts* (pp. 277–280). New York, NY: Peter Lang.

Poor, N., & Skoric, M. M. (2014). Death of a guild, birth of a network: Online community ties within and beyond code. *Games and Culture*, *9*(3), 182–202. https://doi.org/10.1177/1555412014537401

Puente, H., & Tosca, S. (2013). The social dimension of collective storytelling in Skyrim. In *Proceedings of the 2013 DiGRA international conference*.

Puschmann, C. (2019). An end to the wild west of social media research: A response to Axel Bruns. *Information Communication and Society*, *22*(11), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

Raymond, E. S. (1999). *The Cathedral and the Bazaar: Musings on Linux and Open-source by an accidental revolutionary*. Sebastopol, CA: O'Reilly.

Resnick, B. (2016). Researchers just released profile data on 70,000 OkCupid users without permission. *Vox*. Retrieved from https://www.vox.com/2016/5/12/11666116/70000-okcupid-users-data-release

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Sandvig v. Sessions, No. 16-1368 (D.D.C. Mar. 30, 2018).

Smith, B. (2009). FSF settles suit against Cisco. *Free Software Foundation*. Retrieved from https://www.fsf.org/news/2009-05-cisco-settlement.html

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*. https://doi.org/10.1177/0894439319843669

Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks, CA: Sage.

Stricklett, S. (2020). Google v. Oracle: An expansive fair use defense deters investment In original content. *IPWatchdog*. Retrieved from https://www.ipwatchdog.com/2020/01/19/google-v-oracle-expansive-fair-use-defense-deters-investment-original-content/id=117951

Tolkien, J. R. R. (1937). *The Hobbit*. London: George Allen & Unwin.

Tomasello, M. (2010). *Origins of human communication*. Cambridge, MA: MIT Press.

Wellman, B., Boase, J., & Chen, W. (2002). The networked nature of community: Online and offline. *IT & Society*, *1*(1), 151–165.

About the Author

Nathaniel Poor is a Computational Social Scientist who primarily researches online communities, often ones related to games. He is the President and Founder of the Underwood Institute, a non-profit involved in data for good and code for good efforts.