JOURNAL OF DIGITAL SOCIAL RESEARCH

WWW.JDSR.IO ISSN: 2003-1998

VOL. 1, NO. 1, 2019, 17–24

IN SEARCH OF MEANING: WHY WE STILL DON'T KNOW WHAT DIGITAL DATA REPRESENT

Rebekah Tromble*

ABSTRACT

In the early years, researchers greeted the internet and digital data with almost wide-eyed wonder and excitement. The opportunities provided by digital media such as websites, bulletin boards, and blogs—and later by social media platforms and mobile apps—seemed nearly endless, and researchers were suddenly awash in data. The bounty was so great that it required new methods for processing, organizing, and analysis. Yet in all the excitement, it seems that the digital research community largely lost sight of something fundamental: a sense of what all these data actually represent. In this essay, I argue that moving forward, researchers need to take a critical look into, be more open about, and develop better approaches for drawing inferences and larger meaning from digital data. I suggest that we need to more closely interrogate what these data represent in at least two senses: statistical and contextual. In the former instance I call for much greater modesty in digital social research. In the latter, I call for heuristic models that permit bolder, more robust comparisons throughout our work.

*The George Washington University, United States.

In the early years, researchers greeted the internet and digital data with almost wide-eyed wonder and excitement. The opportunities provided by digital media such as websites, bulletin boards, and blogs—and later by social media platforms and mobile apps—seemed nearly endless. Available across geographical distances and, in many instances, not bound by time, it was possible to observe and explore human expression, behavior, connections, and interactions in both new and old forms. And researchers were suddenly awash in data. Indeed, the bounty was so great that it required new methods for processing, organizing, and analyzing the information at hand. "Big data" became the new buzz, with expressions of hope and enthusiasm about all the insights to be gained from its sheer abundance.

Yet in all the excitement, it seems that the digital research community largely lost sight of something fundamental: a sense of what all these data actually represent. In this essay, I argue that moving forward, researchers need to take a critical look into, be more open about, and develop better approaches for drawing inferences and larger meaning from digital data. I suggest that we need to more closely interrogate what these data represent in at least two senses: statistical and contextual.

1 STATISTICAL REPRESENTATIVENESS – CAN WE DRAW UNBIASED INFERENCES FROM OUR SAMPLES?

In statistical terms, we must do a better job of assessing how representative our datasets are of the larger population from which they are drawn. Digital social research often relies on convenience sampling, including when collecting data via platforms' application programming interfaces (APIs). Yet we rarely acknowledge this fact. And we are even less likely to carefully assess the potential implications for our findings.

To illustrate the problem at hand, consider the vast body of research focused on Twitter. Digital researchers have long flocked to this platform. Web of Science identifies 7,343 studies published during the last five years that list "Twitter" as a main topic, and a Google Scholar search for the term "Twitter data" returns more than 6,200 results for 2018 alone. Research on Twitter is so ubiquitous because of the ease of data access. Unlike Facebook, where much of the content is private, the vast majority of tweets are public. And unlike more public spaces such as Instagram and reddit, Twitter offers APIs that provide access to tremendous numbers of posts and their metadata free of charge. Twitter's Streaming API provides access to public tweets in real time, while the Search API provides access to historical tweets.

However, each API carries significant limitations. If a researcher is interested in capturing tweets that match certain keywords, the Search API will only return posts generated within roughly the last week. And according to Twitter's own documentation, the Search API returns tweets based on "relevance," not

"completeness."[1] In other words, the API returns a non-random sample of the full data. The Streaming API, on the other hand, will return the complete population of data matching a keyword query, but only if the matched tweets do not constitute more than 1% of the global volume of tweets at any given moment in time. In other words, the more a keyword is (or set of keywords are) tweeted *or* the fewer tweets are being generated overall (e.g., during major holidays), the more likely one is to get incomplete data. When rate limits are imposed on the Streaming API, the data are truncated. Any tweets above the 1% threshold are simply withheld (Tromble, Storz and Stockmann 2017).

Thus, keyword queries to the Search API virtually ensure non-random data sets, while high-volume captures via the Streaming API are also likely to generate non-random samples. Unfortunately, this means that a great deal of Twitter research is based on statistically biased inferences that in turn undermine conclusions drawn about the social behaviors and relationships under investigation (Tromble, Storz and Stockmann 2017).

One of the only ways to effectively solve this problem for keyword based Twitter research is to purchase the required data. This is not only cost prohibitive, but the solution only applies to data captured in real time. Data purchased from Twitter's historical archive are incomplete; any tweets that have been subsequently deleted or set to private are removed from the archive, and such longitudinal data decay is also non-random (Tromble and Stockmann 2017).

The problem of statistical representativeness is not unique to Twitter. Among platforms that still maintain public APIs, data collection is typically limited to a relatively small number of recent posts or content. Reddit, for instance, permits the capture of just 1,000 posts from a given subreddit. Nor is the problem unique to social media platforms, per se. One common source of digital social research data, the Internet Archive's Wayback Machine, does not contain orphan pages (i.e., those to which no other page links), and websites that include a robot exclusion standard (robots.txt) were long exempted from its crawls.[2]

Of course, not all research will be impacted by such limitations. Projects that collect real-time Twitter data continuously over long periods of time can alleviate the concerns about non-random samples. So too for long-term data collection via reddit's API. And research examining specific websites that have been fully captured by the Wayback Machine will be on firmer ground. But for short-term, snapshot studies that seek some degree of generalizability—whether across websites, platforms, or even *within* a given platform itself—the concerns are substantial. Without better indications of whether and how such data systematically

1 Twitter (nd). Search Tweets, https://developer.twitter.com/en/docs/tweets/search/overview/standard.html, (accessed 11 April 2019).

2 Internet Archive. (nd). Using the Wayback Machine, https://help.archive.org/hc/en-us/articles/360004651732-Using-The-Wayback-Machine (accessed April 11, 2019).

differ from the relevant population, it is difficult to say in statistical terms what our data represent.

2 CONTEXTUAL REPRESENTATIVENESS – WHAT IS THIS AN INSTANCE OF?

In broader contextual terms, digital social research has yet to offer a clear and standard set of heuristics that would facilitate our understanding of how digital data from one platform or space relate to data from another. Though comparative research is increasing (e.g., Boczkowski, Matassi and Michelstein 2018; Bossetta 2018; Rossini et al 2018), studies still tend to be single-platform. In most cases, single-platform studies draw carefully limited conclusions, providing, for example, an analysis of self-presentation on Instagram (Smith and Sanderson 2015) or an exploration of sexualized communication on Snapchat (Charteris and Gregory 2018). In other instances, however, the conclusions are broad and sweeping, suggesting, for example, that we might learn about the impacts of disagreement on social media writ large based on data exclusively from Twitter (Bail et al 2018). In the absence of carefully developed heuristic models or typologies, both approaches miss crucial relationships, context, and, therefore, meaning in the data. In other words, without a better understanding of how data drawn from one digital context relate and compare to data drawn from others, we cannot confidently say what our data are *instances of*.

Consider that vast body of Twitter research again. We turn to Twitter to examine phenomena as disparate as political polarization (Conover et al 2011; Garimella and Weber 2017; Yardi and Boyd 2010), private disclosure (Jin 2013; Walton and Rice 2013), and self-harm (O'Dea et al 2015). But what does it mean in broader terms when we find that political discourse on Twitter is polarized or that it is possible to detect suicide risk factors on the platform? Twitter has relatively few users. In February 2019, the company reported just 126 million daily active users (Twitter 2019) (compared to 1.5 billion for Facebook[3] and 186 million on Snapchat[4]), and users from the United States, who make up the bulk of Twitter engagement, are not representative of the American population in general (Barberá and Rivera 2014). Twitter also has a particular structure and set of design features that shape and constrain communication and social interactions in ways unlike any other social media platform (Bossetta 2018)—let alone the broader digital ecosystem. The variation in such affordances across platforms means that even when actions appear broadly similar—for example, "liking" a post on Twitter vs. Facebook—they may convey very different meanings (Bucher and Helmond 2018). What does political polarization or private disclosure on Twitter tell us about

3 Statista, https://www.statista.com/statistics/346167/facebook-global-dau/ (accessed April 11, 2019).
4 Statista, https://www.statista.com/statistics/545967/snapchat-app-dau/ (accessed April 11, 2019).

political polarization or private disclosure more broadly? Is the answer simply "nothing"?

I do not believe so. However, if we want to make more substantial gains in our understanding of digital social phenomena, we will need to develop—and consistently draw upon—heuristic typologies and models that provide logical guides for comparison and generalization across digital spaces. Such heuristics are incredibly common in other fields. In political science single-country case studies, as well as comparative analyses, are typically guided by foundational typologies that delineate levels and types of democracy (Jaggers and Gurr 1995) or various forms of electoral and party system design (Lijphart 2012). In mass media studies, research is frequently rooted in the typology of "media systems" developed by Hallin and Mancini (2004). These heuristics provide theoretical purchase. We might, for example, expect individual politicians to be more active on social media in majoritarian electoral systems where people directly elect their representatives, as compared to proportional electoral systems where voters select a party, not a specific politician (Tromble 2018). Such heuristic models also strengthen conclusions about the generalizability of our findings. If research finds that individual politicians are more active on social media in one majoritarian electoral system, there is a stronger case for suggesting it likely to be true in others. And a follow-up empirical analysis might directly test this expectation, bolstering cumulative knowledge.

Bossetta's (2018) work on the "digital architectures" of social media offers a potentially valuable starting point for digital social research. Breaking these architectures into four components—network structure, functionality, algorithmic filtering, and datafication—Bossetta demonstrates how these features shape and constrain political campaign communication across Facebook, Twitter, Instagram, and Snapchat in the United States. Such typologies can and should be further developed to include the wide range of digital spaces, not just social media platforms. Architectural features offer one option, but heuristic models can also be rooted in other characteristics of a digital space, including forms of communication (e.g., style, affect, linguistic features) or types of expression (e.g., performative, political). Existing comparative studies naturally draw upon many of these elements already, but without wider-scale and systematic heuristic models, the elements remain disparately engaged, and the larger body of digital social research continues to be disjointed.

3 CONCLUSION

Digital social research is at a crossroads. The heady days of data largess are mostly behind us. Scholars across fields and from different epistemological perspectives seem more willing to acknowledge the limitations we face in our research. My hope is that researchers will take this moment to expand critical reflections and (re)consider how we interpret and derive meaning from digital data.

I have only touched on two forms of data representativeness, statistical and contextual. There are certainly others. But these are important. From a statistical perspective, digital data are rarely perfectly representative. Indeed, given the near impossibility of breaking through proprietary black boxes to examine what data are even available—what in fact *constitutes* the population we might be interested in—demanding such perfection would be pedantic. However, we must be more transparent about the limitations of our data and more cautious in our findings and claims. Here the call is for more modesty. My hope for contextual representativeness, on the other hand, is for greater boldness. Appropriately grounded boldness. But boldness nonetheless. By taking a step back, applying a broader view to the digital landscape as a whole, and developing systematic heuristic models that apply across the ecosystem, digital social researchers will be able to make stronger, more robust claims based on specific empirical data—leading ultimately to a better understanding of what our data represent.

REFERENCES

Barberá, P., and Rivero, G. (2015) 'Understanding the Political Representativeness of Twitter Users', Social Science Computer Review, 33(6), pp. 712–729, https://doi.org/10.1177/0894439314558836

Boczkowski, P. J., Matassi, M., and Mitchelstein, E. (2018) 'How Young Users Deal With Multiple Platforms: The Role of Meaning-Making in Social Media Repertoires', Journal of Computer-Mediated Communication, 23(5), pp. 245–259, https://doi.org/10.1093/jcmc/zmy012

Bossetta, M. (2018) 'The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election', Journalism & Mass Communication Quarterly, 95(2), pp. 471–496, https://doi.org/10.1177/1077699018763307

Bucher, T., and Helmond, A. (2018) 'The Affordances of Social Media Platforms', in Burgess, J., Marwick, A. and Poell, T. (eds), The SAGE Handbook of Social Media, Sage Publications, pp. 233-253, https://doi.org/10.4135/9781473984066.n14.

Charteris, J., and Gregory, S. (2018) 'Snapchat and digitally mediated sexualised communication: ruptures in the school home nexus', Gender and Education, 0(0), pp. 1–17, https://doi.org/10.1080/09540253.2018.1533922

Conover, M. D., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., and Flammini, A. (2011) 'Political Polarization on Twitter', Fifth International AAAI Conference on Weblogs and Social Media. Presented at the Fifth International AAAI Conference on Weblogs and Social Media. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847

Garimella, V. R. K., and Weber, I. (2017) 'A Long-Term Analysis of Polarization on Twitter', Eleventh International AAAI Conference on Web and Social Media. Presented at the Eleventh International AAAI Conference on Web and Social Media. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15592

Hallin, D. C., and Mancini, P. (2004) Comparing media systems: Three models of media and politics. Cambrige: Cambridge university press.

Jaggers, K., and Gurr, T. R. (1995) 'Tracking democracy's third wave with the Polity III data', Journal of peace research, 32(4), pp. 469-482, https://doi.org/10.1177/0022343395032004007.

Jin, S.-A. A. (2013) 'Peeling back the multiple layers of Twitter's private disclosure onion: The roles of virtual identity discrepancy and personality traits in communication privacy management on Twitter', New Media & Society, 15(6), pp. 813–833. https://doi.org/10.1177/1461444812471814

Lijphart, A. (2012) Patterns of democracy: Government forms and performance in thirty-six countries. Second edition. New Haven:Yale University Press.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., and Christensen, H. (2015) 'Detecting suicidality on Twitter', Internet Interventions, 2(2), pp. 183–188, https://doi.org/10.1016/j.invent.2015.03.005

Smith, L. R., and Sanderson, J. (2015) 'I'm Going to Instagram It! An Analysis of Athlete Self-Presentation on Instagram', Journal of Broadcasting & Electronic Media, 59(2), pp. 342–358, https://doi.org/10.1080/08838151.2015.1029125

Rossini, P., Hemsley, J., Tanupabrungsun, S., Zhang, F., and Stromer-Galley, J. (2018) 'Social Media, Opinion Polls, and the Use of Persuasive Messages During the 2016 US Election Primaries', Social Media + Society, 4(3), https://doi.org/10.1177/2056305118784774

Tromble, R. (2018) 'Thanks for (actually) responding! How citizen demand shapes politicians' interactive practices on Twitter', New media & society, 20(2), pp. 676-697, https://doi.org/10.1177/1461444816669158.

Tromble, R., and Stockmann, D. (2017) 'Lost Umbrellas: Bias and the Right to Be Forgotten in Social Media Research', in Zimmer, M. and Kinder-Kurlanda, K. (eds) Internet Research Ethics for the Social Age: New Cases and Challenges. New York: Peter Lang Publishers, pp. 75-91.

Tromble, R., Storz, D., and Stockmann, D. (2017) 'We don't know what we don't know: when and how the use of Twitter's public APIs biases scientific inference', SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3079927.

Twitter (2019, February 7). Q4 and fiscal year 2018 letter to stakeholders. https://s22.q4cdn.com/826641620/files/doc_financials/2018/q4/Q4-2018-Shareholder-Letter.pdf

Walton, S. C., and Rice, R. E. (2013) 'Mediated disclosure on Twitter: The roles of gender and identity in boundary impermeability, valence, disclosure, and stage', Computers in Human Behavior, 29(4), pp. 1465–1474, https://doi.org/10.1016/j.chb.2013.01.033

Yardi, S., and Boyd, D. (2010) 'Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter', Bulletin of Science, Technology & Society, 30(5), pp. 316–327, https://doi.org/10.1177/0270467610380011