Social Science Computer Review
2017, Vol. 35(1) 3-9

DOI: 10.1177/0894439315619589
journals.sagepub.com/home/ssc
SAGE

Citizenship, Social Media, and Big Data: Current and Future Research in the Social Sciences

Homero Gil de Zúñiga[1,2] and Trevor Diehl[1]

Abstract

This special issue of the *Social Science Computer Review* provides a sample of the latest strategies employing large data sets in social media and political communication research. The proliferation of information communication technologies, social media, and the Internet, alongside the ubiquity of high-performance computing and storage technologies, has ushered in the era of computational social science. However, in no way does the use of "big data" represent a standardized area of inquiry in any field. This article briefly summarizes pressing issues when employing big data for political communication research. Major challenges remain to ensure the validity and generalizability of findings. Strong theoretical arguments are still a central part of conducting meaningful research. In addition, ethical practices concerning *how* data are collected remain an area of open discussion. The article surveys studies that offer unique and creative ways to combine methods and introduce new tools while at the same time address some solutions to ethical questions.

Keywords

big data, trace data, ethics of big data, social media, social network analysis, computational social sciences, participation, political discussion, citizenship

It has now been 20 years since Nicolas Negroponte predicted a time when information technologies would permeate every aspect of modern life. In *Being Digital*, Negroponte (1996) envisioned a digital life, where newspapers tailor content to your preferences, touch screens replace mouse pads, and media consumption becomes a highly personalized experience. Information technologies, he argued, would be capable of a subtle, intimate knowledge of individual behavior. Today, news recommender systems, digital social networks, and online shopping malls capture little bits of our daily activities. These bits of data are known as digital trace data (Lazer et al., 2009). Trace data, when collected and stored for research purposes, are often referred to as "big data" (Dumbill, 2012). In just a few decades, computer technologies have accelerated to the point where large, complex data sets have

[1] University of Vienna, Vienna, Austria
[2] Universidad Diego Portales, Santiago, Chile

Corresponding Author:
Homero Gil de Zúñiga, University of Vienna, Wahringer Strasse 29, Vienna, 1090, Austria.
Email: homero.gil.de.zuniga@univie.ac.at

become a standard feature across a range of disciplines. In this sense, Negroponte was correct in his forecast about the integration of communication technologies and daily life. What is less clear, however, is what kind of methods can be used to best tap the knowledge trace data might offer. In other words, how can we interpret and understand human behavior using big data sets?

This question is particularly relevant for social media and political communication scholars. The Internet and social media offer more fluid associations between citizens and the state, groups and political causes, and among individual citizens (Chadwick, 2006). As Bennett and Segerberg (2013) note, new media technologies enable a highly personalized form of political participation. Citizens can contour news flows on social media, complain directly to a municipality on the city Facebook page, mobilize for a political cause on Twitter, or argue about politics on a news comment board. In short, the Internet and social media have altered today's political realm, potentially reinvigorating the way people discuss and express their thoughts about politics, and the way citizens mobilize and engage in political activities (Gil de Zúñiga, 2014, 2015). These changes in communication technologies affect the very nature of how citizens relate to each other and to their governments, shaping a new public sphere (Gil de Zúñiga, 2015; Mourao et al., 2015; Saldaña, McGregror, & Gil de Zúñiga, 2015).

When digital technologies open new spaces for political communication and interaction, these interactions also leave a diverse myriad of data traces in their wake. How to interpret these data remains an enormous challenge. To what extent are social media political interactions representative of underlying social phenomena? One vein of research in this area attempts to use social media data to predict off-line political behaviors. These studies are an attempt to supplement or replace traditional survey metrics (Ecker, this issue; Tumasjan, Sprenger, Sandner, & Welpe, 2010). However, social media use does not always translate to activity outside digital spaces. It is entirely possible, that with big data, we are only looking at the artifacts of one type of user interaction on a particular platform. Another common approach is to equate mentions on social media with some political phenomena. In many of these cases, as Jungherr (2015) correctly points out, correlations between mentions and real-world outcomes (like public opinion or political support) are most likely spurious (see e.g., Metaxas, Mustafaraj, & Gayo-Avello, 2011). The field needs a deliberate attempt to explain the context-specific mechanisms that create patterns in trace data. In addition, the extent to which political activity on social media represents off-line attitudes, behaviors, and processes remains unclear.

In this special issue of *Social Science Computer Review*, large data sets are applied to map political discussion and participation online and, in some cases, also map how these activities connect to off-line political realities. The theoretical and methodological tools applied in this volume are diverse. The differences in approach represent the unsettled nature of using big data sets for social science. This introductory article sets out to highlight some lingering concerns with big data research while showing how the authors in this volume address some of these concerns.

Benefits and Lingering Problems

New methods of inquiry have long been associated with major shifts in the way science is practiced. Telescopes, for example, eventually lead to a reconstruction of how humans think about their place in the universe (Kuhn, 1957). As Kuhn (2012) notes, paradigm shifts do not take place over night or through the work of one person. Instead, new methods of inquiry initially open the door to anomalies. The extent to which the old theories and methods are incapable of explaining new observations determines how, and when, new paradigms arise. Big data sets alone are not enough to represent such a shift in the area of political communication. Sometimes, the term big data is reserved for an area of research dominated by private companies for the purpose of selling products to consumers (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012). One myth in this regard is that a scientific method, driven by causal inference and theoretical questions, is no longer essential. Instead, all

that is necessary is to data mine relationships (Anderson, 2008; Jungherr, 2015; Mayer-Schönberger & Cukier, 2013). On the contrary, in order to better understand the fluid, high-speed world of social media and politics, we still rely on basic principles of social science: validity of constructs, theory that stresses context, generalizability, and ethical concerns.

Thus, it is important to remember what political communication research with big data looks like. Drawing on Shah, Cappella, and Neuman (2015), articles in this issue display the main features of what is referred to as "computational social science," defined by:

> (1) the use of large, complex datasets . . . ; (2) the frequent involvement of "naturally occurring" social and digital media sources and other electronic databases; (3) the use of computational or algorithmic solutions to generate patterns and inferences from these data; and (4) the applicability to social theory in a variety of domains from the study of mass opinion to public health, from examinations of political events to social movements (p. 7).

Computational social science offers many benefits for researchers willing to be creative enough to solve some of the inherent limitations to this approach. First, data collection tools enable users to collect infinitely larger data sets than would ever be imaginable for hand coding. Backstrom, Boldi, Rosa, Ugander, and Vigna (2012) were able to analyze the "degrees of separation" for all Facebook users at the time (721 million users) and their 69 billion friendship links. In another famous online experiment, Bond et al. (2012) discovered that mobilization efforts are more powerful when they spread through an online social network. That study used 61 million Facebook users. These studies trade assumptions about sample size and generalizability for real-world observations of how humans interact in online networks. Although computers aid in the data collection, there may be other labor-intensive considerations. For example, how do we analyze data where statistical models loose power—and becoming less relevant? Similarly, challenges remain in interpreting large numbers and, just as importantly, how to visualize results.

Other questions remain for how to visualize and update results. For example, one study might rely on a visualization (or collection) tool that, due to outside reasons, may no longer be available or no longer valid. In other words, as the technology changes so might decisions about how to collect, visualize, and interpret data, which ultimately affects the means of replicating any given study. And by extension, it may also affect the means of moving social scientific disciplines forward, as it is harder to "stand on the shoulders" of previous research. Thus, the actual practice of big data research in social science is a moving target. Today's representation is different than tomorrows. In short, the findings are also dynamic and in flux. I argue that the discipline needs to be more flexible here. Perhaps one solution is to be more willing to share tools and methods in a repository as a part of the online version of the journal. This way, it may be easier to update findings as technology changes. Unfortunately, this may be impossible, as some scholars have stressed that cost and access to big data tools prevent replication and correction (boyd & Crawford, 2012).

Large sample sizes also raise questions about the generalizability of a particular sample. If the sample size is inclusive of all users on a particular platform (Backstrom, Boldi, Rosa, Ugander, & Vigna, 2012), inferential statistics are not needed, since the entire sample is present. However, large sample sizes of this nature also render hypothesis testing pointless, since any relationship will appear statistically significant (Kramer, Guillory, & Hancock, 2014). On the other hand, even if we get a large sample, more care needs to be taken to compare how any population of social media users (for example Twitter users) might represent the general population. Vargo and Hopp (in this issue) address this concern by purposeful sampling of Twitter users based on data from official population estimates. This type of solution forces researchers to compare data points across multiple data sets.

Often, data sets are so large that manual coding is impossible. Machine learning algorithms are used to do the coding. In these cases, it is important that researchers understand, and at least make an

effort to explain, how a particular algorithm performs on a particular data set. As Hindman (2015) argues, all algorithms may not perform equally. Understanding the features of your data goes a long way to determining the limits and benefits of the particular tool being used. Machine learning can also be applied to smaller data sets (say a subset of the original sample) to test for model fit and offer a check against results gleamed from the whole set.

As with machine learning, certain algorithms may also replace normal methods of data analysis (e.g., Tibshirani, 1996). Instead of a priori hypothesis testing, these tools allow researchers to data mine sources, by finding the most powerful predictors in any given model. In the past, social scientists posed questions, collected data, and tested a set of hypothesis. Big data tools offer researchers a chance to rethink these practices, particularly with regard to over emphasis on p value testing (Hindman, 2015; Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014). However, big data do not relieve the social scientist of their responsibility to ensure the validity of their methods and place findings in an interpretable framework. Many of the articles in this volume address these concerns by combining traditional social science methods with big data. Others make sense of data through the lenses of well-established theories of political reasoning and behavior.

Finally, a lingering concern comes in the form of ethical issues. Most guidelines of ethical research are based on the *Belmont Report*, a guide to ethical treatment of human subjects in medical practice, released in 1978 (Cassell, 2000). The report outlined basic practices of benevolence, respect for persons, and justice. Issues of fair treatment of research subjects in big data research are not easily resolved based on these premises. In general, the field is in desperate need of revised ethical guidelines, particularly in regard to privacy, informational harm, and fair use of publically accessible information (boyd & Crawford, 2012; Fairfield & Shtein, 2014). For example, how do we go about attaining informed consent from millions of Twitter users? What if the data were not initially intended for research purposes? Further, the more an individual leaves behavioral traces in the digital world, the more susceptible they are to informational harm—where any breach of data could result in leaks of potentially sensitive personal information (Fairfield & Shtein, 2014). Researchers should at least consider potential ethical concerns, and when possible, report how they handle such concerns.

Articles in This Volume

Selections in this volume approach the analysis of political digital trace data from a variety of perspectives. The following brief summaries highlight how the authors each tackle some of the limitations of big data research.

Alejandro Ecker (this issue) examines the potential of Twitter data to estimate policy positions at the individual and party level in the Polish *Sejm*—Poland's lower house of parliament. By comparing results with outside data sources, Ecker is able cross-validate results against "real-world" political outcomes. At the aggregate level of party politics, policy positions seem to be well represented on Twitter. However, at the level of the individual legislator, political language used on Twitter does not reliably predict how a particular legislator will vote. Results indicate that cross-validation helps avoid some of generalizability problems inherent in using Twitter data. Such efforts help establish the long path to establishing context-specific criteria for when Twitter feed might represent actual policy preferences and when they do not.

Based on the creation of an original Facebook page app, Chris Wells and Kjersten Thorson (this issue) introduce a creative means of capturing the actual flow of information an individual is exposed to on their Facebook news feed. Social media sites allow users to contour their news feed to their preferences. While many have argued such activity would lead to news avoidance, selective exposure, or even a decrease in inadvertent news exposure (Prior, 2007; Sunstein, 2009), Wells and Thorson are able to empirically asses these curated information flows. The authors ask respondents

to agree to install an app on users Facebook page. When they agree, the app collects data about what a respondent follows, how many friends they have, and what content was in their newsfeed. As the app collects these data, respondents are surveyed about their political behaviors and levels of public affairs knowledge. After the survey is complete, the app is removed from the Facebook account. Big data sets often ignore individual behaviors, while surveys often rely on self-reported measures. Both approaches are limited in their ability to capture the news preferences of individuals on Facebook. By combining methods, the authors are able to verify survey results and rely on a more nuanced analysis of an individual's information environment.

Qinfeng Zhu (this issue) uses Twitter data to track global conversations around the Hong Kong Occupy Central Movement. Using systems theory, Zhu challenges the assumptions that social network spaces are dominated by citizen to citizen, decentralized network activity. Instead, Zhu employs network analysis to argue that conversations about the occupy movement were dominated by elite institutions. In order to better contextualize these findings, the author also used multivariate analysis to account for a region's income and level of political grievances. This additional layer of validation changes the interpretation of the results. Although large institutions dominate the global conversation, a region's level of political grievance was more powerful predictor of network centrality than income.

Kwon and Cho (this issue) question long-held assumptions about civility in public discourse. Some scholars have noted how acidic online conversation spaces can be (Coe, Kenski, & Rains, 2014). However, the authors argue that some swearing is acceptable in political discussion because it raises emotional engagement and draws readers' attention to the conversation. By measuring readers' response to swearing, they are able to challenge the notion that civility is necessary for engagement. Swearing, they argue, might also re-enforce like-minded groups. Although the authors acknowledge that 83,000 comments might not be considered big data, their study is one of the most inclusive for a particular media market (South Korea). In addition, they build on current theories of the role of emotion in political engagement, while using large data sets as an empirical tool—not an end in itself.

Vargo and Hopp also address civility in political discourse. They compare census data with tweets from all 435 congressional districts in the United States. By combining these two different data sources, the authors are able to compare individual demographic and political traits with the type of political discourse in over 400,000 tweets. In addition, the authors took vote outcomes in each district as a measure of polarization. As opposed to conventional thinking on political polarization, those with lower level of polarization engaged in less civil conversations. Triangulating three data sets creates a picture of conversation on twitter that goes far beyond simple pattern recognition.

Brandtzæg (this issue) draws on over 21 million Facebook users to compare gender differences in political participation. In a rare glimpse of global social media habits, Brandtzæg applies developing theories on the role of expression online in spurring civic and political engagement. The author finds that gender gaps and differences in political preferences persist in social media, despite claims that the Internet can act as a gender "equalizer."

Finally, Maireder and colleagues (this issue) introduce two new tools for mapping network connections on twitter, the *audience diversity score* and the *communication connector bridging score*. These measures assess the diversity of a particular actor's followers and highlight the most influential actors in the network, who are potentially able to connect different spheres of communication. In the author's analysis of the conversations on twitter around the Transatlantic Trade and Investment Partnership, they show how and when individuals might be able to spur protest without the help of mainstream media actors.

This volume speaks to the utility of digital trace data for political communication research. Social media and the Internet offer an alternative space to discuss political issues, consume news, and interact with elites—all in pursuit of a more personalized, interactive form of citizenship. As more and

more of our lives moves online, communication scholars have a greater variety of tools and methods available for tracking human behavior than ever before. The era of computational social science is here. However, as Kuhn (2012) notes, the new paradigm will not be set at one point in time. The methods we use today, and more importantly, the theoretical rationale for employing those methods, take time to develop. We are still a long way from consensus in this area. Huge challenges remain, particularly with deriving valid interpretations and treating subjects with care. The articles in this volume represent some of this most forward think in these areas.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, *16*, 108–109. Retrieved from http://www.wired.com/2008/06/pb-theory/

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012, June). Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 33–42). New York, NY: ACM.

Bennett, W. L., & Segerberg, A. (2013). *The logic of connective action: Digital media and the personalization of contentious politics*. Cambridge, MA: Cambridge University Press.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*, 295–298. doi:10.1038/nature11421

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*, 662–679. doi:10.1080/1369118X.2012.678878

Cassell, E. J. (2000). The principles of the Belmont report revisited: How have respect for persons, beneficence, and justice been applied to clinical medicine? *Hastings Center Report*, *30*, 12–21. Retrieved from http://onlinelibrary.wiley.com/doi/10.2307/3527640/pdf

Chadwick, A. (2006). *Internet politics: States, citizens, and new communication technologies*. Oxford, England: Oxford University Press.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*, 658–679. doi:10.1111/jcom.12104

Dumbill, E. (2012). *Planning for big data*. Sebastopol, CA: O'Reilly Media, Inc.

Fairfield, J., & Shtein, H. (2014). Big data, big problems: Emerging issues in the ethics of data science and journalism. *Journal of Mass Media Ethics*, *29*, 38–51. doi:10.1080/08900523.2014.863126

Gil de Zúñiga, H. (Ed.). (2014). *New agendas in communication: New technologies & civic engagement*. New York, NY: Routledge.

Gil de Zúñiga, H. (2015). Toward a European Public Sphere? The Promise and Perils of Modern Democracy in the Age of Digital and Social Media. *International Journal of Communication*. Retrieved from http://ijoc.org/index.php/ijoc/article/view/4783

Hindman, M. (2015). Building better models prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*, 48–62. doi:10.1177/0002716215570279

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*, 235–241. doi:10.1016/j.tics.2014.02.010

Jungherr, A. (2015). *Analyzing political communication with digital trace data*. Cham, Switzerland: Springer.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*, 8788–8790.

Kuhn, T. S. (1957). *The Copernican revolution: Planetary astronomy in the development of western thought* (*Vol. 16*). Cambridge, MA: Harvard University Press.

Kuhn, T. S. (2012). *The structure of scientific revolutions*. Chicago, IL: University of Chicago press.

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

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York, NY: Houghton Mifflin.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data. The management revolution. *Harvard Business Review*, *90*, 61–67. Retrieved from https://hbr.org/

Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011, October). How (not) to predict elections. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 165–171). Los Alamitos, California: IEEE Computer Society – Conference Publishing Services.

Mourao, R. R., Yoo, J., Geise, S., Araiza, J. A., Kilgo, D. K., Chen, V. Y., & Johnson, T. (2015). European Public Sphere| Online News, Social Media and European Union Attitudes: A Multidimensional Analysis. *International Journal of Communication*. Retrieved from http://ijoc.org/index.php/ijoc/article/view/2990

Negroponte, N. (1996). *Being digital*. New York, NY: Vintage Books.

Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge, England: Cambridge University Press

Saldaña, M., McGregror, S., & Gil de Zúñiga, H. (2015). Social media as a public space for politics: Cross-national comparison of news consumption and participatory behaviors in the United States and the United Kingdom. *International Journal of Communication*. Retrieved from http://ijoc.org/index.php/ijoc/article/view/3238

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*, 6–13. doi:10.1177/0002716215572084

Sunstein, C. R. (2009). *Republic. com 2.0*. Princeton, NJ: Princeton University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288. Retrieved from http://www.jstor.org/stable/2346178

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *International Conference on Weblogs and Social Media*, *10*, 178–185. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852Predicting

Author Biographies

Homero Gil de Zúñiga received his PhD in politics at Universidad Europea de Madrid and PhD in mass communication at the University of Wisconsin–Madison. He holds the Medienwandel Professorship at the University of Vienna, where he directs the Media Innovation Lab (MiLab). He also serves as a research fellow at Universidad Diego Portales, Chile and Research Associate at Princeton University. His research addresses the influence of new technologies and digital media over people's daily lives, as well as the effect of such use on the overall democratic process. Email: homero.gil.de.zuniga@univie.ac.at

Trevor Diehl, MA, is a doctoral student and Research Assistant at the Department of Communication at the University of Vienna, Austria. His research interests include social media and politics, science communication, and journalism practice.