Selected Papers of #AoIR2018:
The 19th Annual Conference of the
Association of Internet Researchers
Montréal, Canada / 10-13 October 2018

A DIVISION OF LABOR: THE ROLE OF BIG DATA ANALYSIS IN THE REPERTOIRE OF INTERNET RESEARCH METHODS

In recent years, large-scale analysis of log data from digital devices – often termed "big data analysis" (Lazer, Kennedy, King, & Vespignani, 2014) – have taken hold in the field of internet research. Through Application Programming Interfaces (APIs) and commercial measurement, scholars have been able to analyze social media users (Freelon 2014) and web audiences (Taneja, 2016) on an uprecedented scale. And by developing digital research tools, scholars have been able to track individuals across websites (Menchen-Trevino, 2013) and mobile applications (Ørmen & Thorhauge 2015) in greater detail than ever before. Big data analysis holds unique potential for studying communication in depth and across many individuals (see e.g. Boase & Ling, 2013; Prior, 2013).

At the same time, this approach introduces new methodological challenges in the transparency of data collection (Webster, 2014), sampling of participants and validity of conclusions (Rieder, Abdulla, Poell, Woltering, & Zack, 2015). Firstly, data aggregation is typically designed for commercial rather than academic purposes. The type of data included as well as how it is presented depend in large part on the business interests of measurement and advertisement companies (Webster, 2014). Secondly, when relying on this kind of secondary data it can be difficult to validate the output or techniques used to generate the data (Rieder, Abdulla, Poell, Woltering, & Zack, 2015). Thirdly, often the unit of analysis is media-centric, taking specific websites or social network pages as the empirical basis instead of individual users (Taneja, 2016). This makes it hard to untangle the behavior of real-world users from the aggregate trends. Lastly, variations in what users do might be so large that it is necessary to move from the aggregate to smaller groups of users to make meaningful inferences (Welles, 2014). Internet research is thus faced with a new research approach in big data analysis with potentials and perils that need to be discussed in combination with traditional approaches.

This panel explores the role of big data analysis in relation to the wider repertoire of methods in internet research. The panel comprises four presentations that each sheds light on the complementarity of big data analysis with more traditional qualitative and quantitative methods.

The first presentation opens the discussion with an overview of strategies for combining digital traces and commercial audience data with qualitative interviews and quantitative survey methods. The next presentation explores the potential of trace data to improve upon the experimental method. Researcher-collected data enables scholars to operate in a real-world setting, in contrast to a research lab, while obtaining informed consent from participants. The third presentation argues that large-scale audience data provide a unique perspective on internet use. By integrating census-level information about users with detailed traces of their behavior across websites, commercial audience data combines the strength of surveys and digital trace data respectively. Lastly, the fourth presentation shows how multi-institutional collaboration makes it possible do document social media activity (on Twitter) for a whole country (Australia) in a comprehensive manner. A feat not possible through other methods on a similar scale. Through these four presentations, the panel aims to situate big data analysis in the broader repertoire of internet research methods.

References

Boase J and Ling R. (2013) Measuring Mobile Phone Use: Self-Report Versus Log Data. *Journal of Computer-Mediated Communication* 18: 508-519.

Freelon D. (2014) Tweeting to Power: The Social Media Revolution in American Politics. *Political Communication* 31: 502-505.

Lazer D, Kennedy R, King G and Vespignani A. (2014) The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343: 1203-1205.

Menchen-Trevino E. (2013) Collecting vertical trace data: Big possibilities and big challenges for multi-method research. *Policy and Internet* 5: 328–339.

Prior M. (2013) The Challenge of Measuring Media Exposure: Reply to Dilliplane, Goldman, and Mutz. *Political Communication* 30: 620-634.

Rieder B, Abdulla R, Poell T, Woltering R and Zack L. (2015) Data critique and analytical opportunities for very large Facebook Pages: Lessons learned from exploring "We are all Khaled Said". *Big Data & Society* 2.

Taneja H and Webster JG. (2016) How Do Global Audiences Take Shape? The Role of Institutions and Culture in Patterns of Web Use. *Journal of Communication* 66: 161-182.

Webster JG. (2014) *The marketplace of attention. How audiences take shape in a digital age,* London and Cambridge, MA: MIT Press.

Welles BF. (2014) On minorities and outliers: The case for making Big Data small. *Big Data & Society* 1: 1-2.

Ørmen J and Thorhauge AM. (2015) Smartphone log data in a qualitative perspective. *Mobile Media & Communication* 3: 335-350.

COMBINING DIGITAL TRACE DATA WITH RESEARCH METHODS ON A GLOBAL SCALE

Rasmus Helles
University of Copenhagen

Jacob Ørmen
University of Copenhagen

Signe Sophus Lai
University of Copenhagen

Klaus Bruhn Jensen
University of Copenhagen

This paper looks at the role big data can play in comparative media research on a global scale. It does so with reference to an ongoing research project, The Peoples' Internet (PIN), which studies internet use in three world regions. In this way, the project functions as a case for mapping out the unique contribution of big data analysis in relation to traditional methods of internet research.

The PIN project

The PIN project (http://peoplesinternet.ku.dk/) investigates how ordinary people use the internet in daily life across 5 European countries, the US, and China. It relies on a mixed-methods approach combining ethnography with large-scale surveys, document analysis and big data analysis. The ethnography component will be carried out as 6-month field work studies (relying primarily on interviews and observations) in China, the US and Denmark. Simultaneously, population surveys in each participating country will map out internet use practices in relation to communication patterns more broadly (Jensen & Helles, 2011). These methods are supplemented by document analysis of legal and policy texts to establish the political and economic context for the internet in each country under study.

In this paper, we focus on the big data component of the project. The big data in PIN will be digital trace data (Freelon, 2014) of online users collected by the audience measurement company ComScore. This data makes it possible to analyze audiences for various web sites (and possibly mobile applications) as well as *user flows* (Jensen, 2012) across the web. In this way, digital trace data provide a lens on internet usage across world regions that complement the ethnographies and survey research. This paper maps out the various roles digital trace data play in relation to the other methods in the project.

Strategies for mixing digital trace data with other methods

The methodological framework of PIN is to collect data concurrently across methods and analyze data separately as well as in combination. Accordingly, all methods applied play an equal part in the analysis. At the same time, each of the methods provides a unique perspective on internet use, which makes it necessary to both compare and contrast findings across methods. To do this, the project relies on various strategies for mixing methods (inspired by Greene, Caracelli, & Graham, 1989).

Complementarity. A prime reason for combining methods is to research several aspects of the phenomena under study which no single method can cover in itself (Greene et al., 1989). For instance, research has found behavioral trace data to be more suitable for capturing precise communication (Boase & Ling, 2013) or media use (Prior, 2009) patterns than surveys. In the PIN context, the ComScore data is able to map out the popularity and audience overlap for various web genres (such as news sites, social network sites, video streaming sites) on a more detailed level than is feasible through other methods. Conversely, the ComScore data cannot grasp important aspects of internet use such as what happens within the different constituents of social network sites or in mobile apps. The surveys and ethnographies will have to fill the void here.

Corroboration. Another goal of the project is to establish confidence in the empirical findings. This strategy is often termed *triangulation* (for overview see Greene et al., 1989). Although originally used as a technique to conduct precise geographical measurement, triangulation is now applied in a more metaphorical sense in social research to corroborate findings through different methods (Blaikie, 1991). By seeking corroboration, it is possible to lend credibility to the overall research design. In PIN, the internet habits of individual participants, as located through ethnographies, can be compared to the audience patterns indicated by digital trace data. If these findings correspond to each other, it makes a stronger case for trusting data sources, the sampling techniques used as well as the interpretations of findings.

Contestation. Although often overlooked, an important aspect of combining methods is to look for deviations and contradictions. As well as establishing agreement, findings from various methods might also point in vastly different directions. Qualitative interviews might paint a different picture of media-related behavior than surveys or digital trace data. In this way, the methods in combination can expose contradictions such as the well-known “paradox of popularity” (Meijer, 2007), whereby people tend to express contempt for some genre of media content (in interviews) they nonetheless spend a lot of time engaging with (as elucidated by e.g. digital trace data). This should not be interpreted as a sign of incompatibility between methods, but rather as a core strength of mixing methods. These contradictions are vital for challenging existing theories and pushing new ones forward.

Enhancement: Last but not least, we envision that the digital trace data can be used to enhance the value of other data sources. One way to do this would be to enrich one data source with another (Salganik, 2016), e.g. by estimating the likelihood that respondents in the survey would showcase the behavior observed in the digital trace data (based on matching socio-demographic characteristics available in both types of data). Another way would be to augment data sources with each other (Boase, 2016),

e.g. by tracing the online behavior of participants through their digital devices in the ethnographies and integrate these data in subsequent interviews.

Summing up, the role of big data in the PIN project is to supplement, underpin and contrast the other methods used. Importantly, big data does offer unique perspectives on internet use that are hard to cover with other methods. Whereas digital trace data are more suitable for looking into specific uses of the web (in particular types of websites used) on a national and global scale, it cannot provide an encompassing picture of internet use. For this, we still need traditional methods for internet research such as quantitative surveys and qualitative ethnographies.

References

Blaikie, N. (1991). A critique of the use of triangulation in social research. *International Journal of Methodology, 25*(2), 115-136. doi:10.1007/BF00145701

Boase, J. (2016). Augmenting Survey and Experimental Designs with Digital Trace Data. *Communication Methods and Measures, 10*(2-3), 165-166. doi:10.1080/19312458.2016.1150975

Boase, J., & Ling, R. (2013). Measuring Mobile Phone Use: Self-Report Versus Log Data. *Journal of Computer Mediated Communication, 18*(4), 508-519. doi:10.1111/jcc4.12021

Freelon, D. (2014). On the Interpretation of Digital Trace Data in Communication and Social Computing Research. *Journal of Broadcasting & Electronic Media, 58*(1), 59-75. doi:10.1080/08838151.2013.875018

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a Conceptual Framework for Mixed-Method Evaluation Designs. *Educational evaluation and policy analysis, 11*(3), 255-274. doi:doi:10.3102/01623737011003255

Jensen, K. B. (2012). Communication in Context. In K. B. Jensen (Ed.), *A Handbook of Media and Communication Research* (2 ed.). London & New York, NY: Routledge.

Jensen, K. B., & Helles, R. (2011). The internet as a cultural forum: Implications for research. *New Media & Society, 13*(4), 517-533. doi:10.1177/1461444810373531

Meijer, I. C. (2007). The Paradox of Popularity. *Journalism Studies, 8*(1), 96-116. doi:10.1080/14616700601056874

Prior, M. (2009). The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure. *Public Opinion Quarterly, 73*(1), 130-143. doi:10.1093/poq/nfp002

Salganik, M. J. (2016). *Bit By Bit: Social Research in the Digital Age*. Open Review. Retrieved from http://www.bitbybitbook.com/

TRACE PARTNERSHIP: COLLECTING REAL-WORLD BEHAVIORS AND SELF REPORTS

Ericka Menchen-Trevino
American University

Many of the most persistent and difficult questions in the social sciences involve relating people's behaviors to their attitudes. One such issue that has been the focus of my research is selective exposure to political communication. That is, to what extent do people select attitude-consistent information and avoid attitude-dissonant information about politics or policy? At issue is the relationship between attitudes in the mind (political or policy views) and behaviors (selecting or avoiding information). This may seem like an un-controversial topic, particularly in the Trump era. However, there have been decades-long controversies about the extent and nature of selective exposure (Feldman, Stroud, Bimber, & Wojcieszak, 2013; Dvir-Gvirsman, Tsfati, & Menchen-Trevino, 2014; Hart et al., 2009; Sears & Freedman, 1967).

There are two established social science methods that typically collect self-reported and behavioral data, albeit in very different contexts, experiments and ethnography. In experiments and ethnography behavior is observed and recorded, traditionally by humans but also potentially by machines. Self-reports are typically collected by questionnaires in experiments and with in-depth interviews in ethnography. Here I am arguing for a third approach in the digital era, partnering with participants to examine real-world digital trace behavior and self-reported attitudes. Digital traces are records of behavior recorded by digital technologies, such as web browsing history, phone location history, and social media activity logs. Often, these traces are accessed by researchers through agreements with or application programming interfaces (APIs) provided by the companies that generate them, e.g. the New York Times website, cell-phone service providers, Twitter or Facebook. This platform-centric approach to trace data gathering is useful, but limited to a single platform, and ethical concerns about privacy appropriately prevent sharing the identifying information that would be needed to recruit individuals to participate in other studies. By partnering with users who have access to their own traces, researchers can gather cross-platform trace data and link it with self-reported data such as survey responses or interviews. I will call this approach *trace partnership*. I used this approach in my work where I collected digital traces, survey responses, and in-depth interviews with participants (Menchen-Trevino, 2012; Menchen-Trevino & Karr, 2012). This is similar to what I have discussed as the collection of vertical trace data, but focuses on the research method rather than the data (Menchen-Trevino, 2013).

Several other projects and research groups, discussed in detail in the full paper, use trace partnership as a practical way to advance various fields. Labeling trace partnership as a methodological approach provides a conceptual toolkit for researchers to engage across methodological divides and recognize common challenges and opportunities toward the goal of strengthening social research in the digital era.

Methodological Distinctions

Although the distinction between quantitative and qualitative methods is important to describing contemporary research traditions (see Goertz & Mahoney, 2012) a more useful distinction in the context of digital trace data is to differentiate inductive versus deductive approaches. Experiments are necessarily deductive, with pre-defined variables and methods resulting in a support or lack of support for a hypothesis. Although an ethnographer enters the field with particular interests, they often shift their inquiry based on conditions in the field, using an inductive approach.

Trace datasets, which often contain textual data like tweets are frequently so large that reading them is impossible. Any overview of a large dataset is necessarily quantitative, but not necessarily deductive. Some quantitative methods like factor analysis, network analysis, and machine learning are primarily inductive.

Focusing on combining digital traces and self-reports, therefore, does not necessitate quantitative or qualitative research alone. It does, however, involve an inductive approach, particularly as relating traces to self-reports involves a process of triangulation and identification of complementarities (Blok & Pedersen, 2014). To the extent that the digital traces collected for a project are too numerous to examine with qualitative methods alone, and this is usually the case, a trace partnership does require quantitative methods. While quantitative inductive work is necessary, it could be part of an otherwise qualitative approach, e.g. incorporating web browsing history log analysis into an ethnography of digital journalists, or a quantitative approach, such as a survey of internet users.

Partnership Through Informed Consent

It is essential that social researchers distinguish themselves from corporations and governments plying uninformed consent, or no consent at all, from individuals to collect their digital traces. This means that researchers need to be public educators, and risk scaring away some participants by educating them about what the data they are asking them for can actually contains. This is the only long-term sustainable path forward for social researchers collecting digital traces, and allies researchers with the public interest in the right to access their own data. The full paper offers practical information visualization techniques for better informing the consent process for digital traces to make the partnership between users and researchers meaningful.

References

Blok, A., & Pedersen, M. A. (2014). Complementary social science? Quali-quantitative experiments in a Big Data world. *Big Data & Society*, *1*(2), 2053951714543908. https://doi.org/10.1177/2053951714543908

Dvir-Gvirsman, S., Tsfati, Y., & Menchen-Trevino, E. (2014). The extent and nature of ideological selective exposure online: Combining survey responses with actual web log data from the 2013 Israeli Elections. *New Media & Society*. https://doi.org/10.1177/1461444814549041

Feldman, L., Stroud, N. J., Bimber, B., & Wojcieszak, M. (2013). Assessing Selective Exposure in Experiments: The Implications of Different Methodological Choices. *Communication Methods and Measures*, *7*(3-4), 172–194. https://doi.org/10.1080/19312458.2013.813923

Goertz, G., & Mahoney, J. (2012). *A Tale of Two Cultures : Qualitative and Quantitative Research in the Social Sciences*. Princeton, N.J.: Princeton University Press.

Hart, W., Albarrac n, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555–588.

Menchen-Trevino, E. (2012, December). *Partisans and Dropouts?: News Filtering in the Contemporary Media Environment*. Northwestern University, Evanston, Illinois.

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

Menchen-Trevino, E., & Karr, C. (2012). Researching real-world Web use with Roxy: Collecting observational Web data with informed consent. *Journal of Information Technology & Politics*, *9*(3), 254–68. https://doi.org/10.1080/19331681.2012.664966

Sears, D. O., & Freedman, J. L. (1967). Selective exposure to information: A critical review. *Public Opinion Quarterly*, *31*(2), 194–213.

THEORETICAL PURCHASE FROM "AUDIENCE CENTRIC LOGS"

Harsh Taneja
University of Illinois Urbana-Champaign

Angela Xiao Wu
New York University

In this abstract, I introduce "audience centric log data," a class of log data provided by commercial audience measurement companies. First, I acknowledge the turn in Internet research towards "log data" and note some of their limitations. Then I focus on audience centric log data and demonstrate how they combine essential characteristics of both surveys and log data. I conclude with an overview of some specific usage measures that these data capture and how through relational techniques such as social network analysis scholars can harness these to advance Internet research. The presentation will elaborate on ideas discussed in this note with examples from empirical studies conducted by the author(s) and their collaborators in the last five years.

"Big Data" now attract significant scholarly interest in both social sciences and humanities (boyd & Crawford, 2012). Internet research is no exception. Recognizing the inherent difficulties of survey research to accurately gauge people's exposure to content in a high choice media environment, Internet researchers take immense interest in "log data" – defined loosely as collections of traces from users' online activities. Wikipedia editing histories, repositories of Tweets by Twitter handles, server side "web analytics", are all examples of such logs. The resulting "big data" are large-scale census level estimates of digital media use. Their popularity across fields has germinated a new academic discipline termed "computational social science".

Log data have attracted fair share of criticism both from substantive and ethical standpoints. Substantively, data from server logs (such as web analytics) or from virtual profiles (e.g., Twitter, Wikipedia) does not have information about actual user characteristics and attitudes. Further, they rarely represent any meaningful target population. Finally, analyzing user behavior on a single domain such as one news website, or one social media site such as Twitter or Wikipedia ignores and obscures their usage of other digital media outlets. Ethically, one may argue that these data are analyzed without informed consent.

Yet there is a special class of "log data" free of these limitations. These are "audience measurement (panel) data", generally collected by commercial market research firms at the behest of media companies and advertisers. First collected for radio and television, these combine essential characteristics of both survey and log data. Common examples include *Nielsen Ratings* and *comScore media metrix*. Collecting such data is a multi-step process (Webster, Phalen & Lichty, 2014). First through a nationally representative sample is selected based on a large- scale study such as the population census. The selected people (or households) are carefully profiled on a range of demographic and psychographic characteristics. A "meter" then logs usage for the chosen panel of respondents directly through the device. Therefore these data are sometimes referred to as "audience centric" log data (to differentiate them "server centric" log data). Hence

with audience centric log data, researchers potentially have information about audience characteristics, and data about their usage across a range of outlets. Further these panels are quite representative of the target populations whose usage they intend to measure, and people need to consent to be on such panels. Traditionally limited to specific mediums (such as radio, television, computers or mobiles) audience measurement firms are increasingly trying to measure cross-platform media use. It is also worth noting that media industry has collected and worked with metered "audience centric log data" for at least six decades, much before social science researchers became interested in "big data".

Despite their promise, audience centric log data have found limited use among academics. Traditionally research in certain subfields of media communication has been highly critical and even cynical of audience measurement data. Further most communication departments lacked the computational resources to deal with such voluminous logs of data. Surprisingly this reluctance continues among Internet researchers, many of whom possess the requisite training and infrastructures to handle log data. Part of this reluctance comes from the proximity of computational social scientists to the Silicon Valley, which has successfully propagated among Internet researchers an imaginary of the "Internet as anti –television" (Sandvig, 2015) and in this perspective commercial audience measurement is most strongly associated with Nielsen television ratings. Thus for a typical computational social scientist while Wikipedia or Twitter are cool digital laboratories for advancing social science, audience measurement panels come loaded with the unfavorable baggage of their traditional media legacies.

A second technical limitation is that often providers make audience centric log data available to researchers in the aggregate at the level of media outlets rather than at the individual user level. Traditionally, social science theories have been built with data collected at the individual level (such as in surveys). Although it is hard to alter the ideological disinclination of certain academics to use these data, one can certainly address the technical limitations, which I will discuss in what remains.

Specifically, although audience centric log data are generally provided in the aggregate, these firms also provide "relational" data. That is how do people use media outlets in relation to one another. First audience measurement firms provide a measure called "audience duplication" – which is the extent of overlap between two media outlets or in other words, the proportion of users of a given outlet who also use another outlet. These pairwise audience duplication data can be used to create an audience duplication matrix, which effectively is a network of media outlets connected to each other through shared audience traffic. A second measure of interest these panels provide is "user clickstreams" which essentially capture user's temporal order of using various outlets. Thus clickstream data informs us that of all people who visited a given website A, how many of them were on another given website immediately before and after visiting A. These data can also be converted to pairwise "audience flow" matrices and unlike audience duplication matrices they also have directionality, which allows for rich analysis of not only people's shared visitation patterns but their order of visiting. Clickstream data for example can help discern the role of search engines and social networks in influencing specific patterns of news usage.

Finally, I recognize that bulk of social science theories relate behavior to user characteristics such as demographics. Although reporting information in the aggregate, audience centric log data do provide demographic breakouts of whatever information they report on - at least on key variables such as gender, age, household income and race. Thus the matrices based on relational measures such as duplication and clickstreams I just described can be obtained for the same media outlet for different demographic segments and usage can be compared between them.

In sum, I hope that by elaborating on the nature of and possibilities with audience centric log data, this presentation will convince more Internet researchers to embrace this class of log data.

References

boyd, danah, & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA. *Information, Communication & Society*, *15*(5), 662–679.

Sandvig, C. (2015). The Internet as the Anti-Television: Distribution Infrastructure as Culture and Power. In: L. Parks and N. Starosielski (eds.), *Signal Traffic: Critical Studies of Media Infrastructures*, pp. 225-245. Chicago: University of Illinois Press.

Webster, J. G., Phalen, P. F., & Licthy, L. W. (2014). *Ratings analysis: Audience measurement and analytics*. New York: Routledge.

A MULTI-INSTITUTIONAL APPROACH TO 'BIG SOCIAL DATA': THE TrISMA PROJECT

Axel Bruns
Queensland University of Technology

Introduction

Methods for the gathering and analysis of large datasets about communicative interactions between users, especially on digital and social media platforms, have become increasingly prominent in the field of Internet research in recent years. This is sometimes aligned with a push towards more quantitative perspectives in communication research, but often also enables the development of new mixed-methods approaches where quantitative analytics for large datasets are used to pinpoint subsets of the data that would benefit especially strongly from a further, detailed qualitative exploration and assessment, for instance through close reading and manual coding approaches.

Any quantitative, qualitative, or mixed-methods analyses that draw on such 'big social data' are necessarily always limited by the quality and reliability of the datasets underlying them, however. Alongside the rise of 'big data' in Internet research we have therefore also seen the emergence of a body of literature that critically reviews the limitations of 'big data' as an overall concept, and of specific sources of 'big data' as they are commonly used in the field. These include overall challenges such as boyd & Crawford's influential "provocations" about 'big data' (2012), as well as detailed analyses especially of the limitations, reliability, and representativeness of the various sources of Twitter data that are particularly widely used in recent scholarship (e.g. Gerlitz & Rieder, 2013; Driscoll & Walker 2014; Bruns & Burgess 2015; Weltevrede, 2016).

Limitations of Current 'Big Data' Approaches

In the field of social media research, such studies have shown that much current scholarship, even when it works with very large datasets, continues to work with data that are subject to a range of severe limitations. Common Twitter data gathering techniques, for instance, continue to rely largely on the tracking of sets of hashtags and/or keywords; although these can generate some very large datasets (comprised of millions or tens of millions of tweets), they nonetheless miss out on important aspects of the communicative process that would be valuable for the full analysis of specific practices, issues, or events: for instance, such hashtag datasets do not contain any of the tweets preceding or responding to a matching tweet unless those tweets themselves also contain the same hashtag. Working with these datasets is analogous to listening in on only one side of a multi-sided phone conversation, therefore, and complicates or prevents any research approaches that seek to examine the full conversations.

Similarly, data gathering approaches that proceed in this way from a set of search terms fundamentally lack context; while it is possible to establish patterns in a given dataset,

and compare them against other, similar datasets, there is usually no baseline information on total platform activity against which they might be benchmarked. A major event (a celebrity death, a political scandal) can be assessed by determining the number of hashtagged tweets it generates, for instance – but how does this number compare to the total volume of tweets posted to Twitter during the same timeframe? More specifically, how many of these tweets were posted by users in a given demographic category, or from a specific geographic region?

Towards More Comprehensive 'Big Social Data' Infrastructures

Some such data may be available from platform providers or their third-party data resellers. In theory, researchers could pay for access to Twitter's global 'firehose' of all tweets, or to equivalent datasets from other platform providers, but both the costs and the infrastructure required to ingest and store such vast quantities of data are likely to be insurmountable hurdles for most individual projects. This paper outlines one possible solution to this problem (as well as the potential pitfalls with this approach): the formation of multi-institutional consortia to underwrite the development and operation of the next generation of 'big social data' infrastructure. It focusses on the TrISMA: Tracking Infrastructure for Social Media Analysis (Bruns *et al*. 2016) project, supported by seven Australian universities, the National Library of Australia, and the Australian Research Council.

TrISMA provides the infrastructure to gather data from a number of leading social media platforms. On Twitter, for instance, it has identified some four million Australian accounts from a global userbase of 1.4 billion accounts (as of early 2016), and mapped the follower relations amongst them; it gathers new public tweets from these accounts on a continuing basis (capturing an average of 1.3 million new tweets per day). In the absence of country-specific 'firehose' offerings from Twitter or its data resellers, this dataset represents the closest available equivalent to an Australian 'firehose'; it constitutes a comprehensive repository of Australian Twitter activity independent of predetermined keywords, hashtags, or other features, and offers a reliable baseline for the overall volume of domestic Twitter activity.

The Challenges of Multi-Institutional Data Infrastructures

The deployment and use of this shared infrastructure also presents unique new challenges to the developers and researchers involved, however. First, the technical challenges inherent in gathering, processing, and storing such large datasets (the Twitter collection alone now contains more than 2.2 billion tweets) are significant, and the changeable nature, limited documentation, and vague Terms of Service of the Twitter Application Programming Interface complicate this further. Second, the multi-institutional nature of the project introduces coordination challenges: for instance, while researchers at all member institutions are able to access the infrastructure, it is necessary to ensure that they have also received the required ethics clearances and methods training before they do so. Third, the social media analytics methods required to use TrISMA data remain emergent and experimental, and rely on a number of key data processing tools and skills; a pronounced need for coordinated research training for users of the infrastructure has therefore also become apparent. Finally, with an

increasing number of projects drawing on the infrastructure, ensuring its stability and reliability has also become a crucial concern.

This paper reviews these challenges and presents some of the solutions emerging. It also outlines the unique contributions that this multi-institutional 'big social data' infrastructure is able to make to the field, over and above more limited data gathering frameworks. Amongst these are, for Twitter, the ability to work with a comprehensive dataset of domestic Australian tweets; to trace more complex communicative exchanges independent of the keywords and hashtags in each contributing tweet; and to examine the intersections between communicative activity (tweets) and underlying structural factors (follower relations). It closes by outlining further needs in infrastructure and methods development.

References

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–79. doi:10.1080/1369118X.2012.678878.

Bruns, Axel, Burgess, Jean, Banks, John, Tjondronegoro, Dian, Dreiling, Alexander, Hartley, John, Leaver, Tama, Aly, Anne, Highfield, Tim, Wilken, Rowan, Rennie, Ellie, Lusher, Dean, Allen, Matthew, Marshall, David, Demetrious, Kristin, and Sadkowsky, Troy. (2016). *TrISMA: Tracking Infrastructure for Social Media Analysis*, http://trisma.org/.

Burgess, Jean, and Axel Bruns. 2015. "Easy Data, Hard Data: The Politics and Pragmatics of Twitter Research after the Computational Turn." In *Compromised Data: From Social Media to Big Data*, edited by Ganaele Langlois, Joanna Redden, and Greg Elmer, 93–111. New York: Bloomsbury Academic.

Driscoll, Kevin, and Shawn Walker. 2014. "Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data." *International Journal of Communication* 8 (0): 20. http://ijoc.org/index.php/ijoc/article/view/2171.

Gerlitz, Carolin, and Bernhard Rieder. 2013. "Mining One Percent of Twitter: Collections, Baselines, Sampling." *M/C Journal* 16 (2). http://journal.media-culture.org.au/index.php/mcjournal/article/view/620.

Weltevrede, Esther. 2016. "Repurposing Digital Methods: The Research Affordances of Platforms and Engines." PhD, Amsterdam: University of Amsterdam.