Studying Big Data – ethical and methodological considerations

Anders Olof Larsson

Department of Media and Communication,
University of Oslo

a.o.larsson@media.uio.no

Introduction

New technologies for communication tend to raise certain expectations regarding the more or less overwhelming societal influence arising from them. Such was the case with radio and television – and consequently, also during the mid-1990s, when the Internet started to grow in popularity throughout much of the Western world. While more traditional or established forms of media remain a major part in our everyday media diets, the Internet has indeed come to play an important role in our day-to-day activities. Needless to say, such a move to a variety of online environments is of significant interest to a variety of scholars from the social sciences and the humanities. This chapter presents an overview of some of the challenges of performing research on the activities taking place in such environments. While my personal experience with this type of research is geared more towards perspectives often

associated with the social sciences – specifically various aspects of online political communication – it is my hope that the concerns raised will also resonate with readers who approach studies of online environments from other perspectives.

Specifically, the focus here is on the phase in the development of the Internet often referred to as the «Web 2.0.» While there is not detailed agreement on what this supposed second stage of the World Wide Web entails, attempts towards a definition tend to revolve around ideas of increased user participation (e.g. O'Reilly, 2005). As suggested by Small, «Whereas Web 1.0 was 'read-only,' Web 2.0 is 'read/write,' allowing online users to contribute to the content rather than just being passive viewers» (Small, 2012: 91). Often discussed in conjunction with so-called social media services (such as Twitter or Facebook), services that are more or less dependent on such active user communities, the 2.0 variety of the Internet has received plenty of societal as well as scholarly interest – as well as its fair share of what must be labeled «hype.» The uses of such services, then, are often thought to yield «Big Data» – orderly traces of online activity of potential interest to researchers in multiple fields. While usage rates and modes of social media vary, these types of services are arguably here to stay – although we should not expect the services currently in fashion to remain so forever. What we can expect is for the data deluge created by these services to persist – and to grow in size.

While the term «Big Data» can be tagged onto a multitude of discussions regarding the increased possibilities of tracing, archiving, storing and analyzing online data, the specific appropriation of the term here deals with how masses of data are gathered from social media services like the ones discussed previously and subsequently analyzed for research purposes. In so doing, I would like to discuss two broad thematic groups of challenges that researchers often face when doing research on social media. The first group deals with

ethical issues, while the latter concerns more methodological possibilities and problems. Before delving into these issues, though, we need to look a bit closer at the term «Big Data» and its many connotations.

Big Data – size is everything?

As with the Web 2.0 concept, the term «Big Data» carries with it a number of differently ascribed meanings and ideas. As the name implies, definitions often involve discussions regarding the swelling size of the data sets that researchers as well as other professionals now have to deal with. Indeed, the growing use of social media combined with the increased sophistication of tools for «scraping» such online environments for data has provided «an ocean of data» (Lewis, Zamith, and Hermida, 2013: 35) that mirrors such new activities: Facebook updates, tweets, Instagram photos, etc. Such vast amounts of data can be collected and curated from a number of different services and with several purposes in mind – for scholarly efforts, this has led to claims like Chris Anderson's suggestion that Big Data could lead to «the end of theory» (Anderson, 2008). In essence, the quantities of data now readily available, supposedly at the click of a button, could render scholarly practices like employing theory and sampling rationales obsolete. While Anderson might be correct in that approaches to sampling and data collection more generally when it comes to research dealing with the online environment need to be revisited and reformulated in some instances, the argument is made here that no matter the vastness of the data, the need for some form of theoretical rationale in order to separate «noise from signal» (González-Bailón, 2013: 154) is evident. Indeed, searching for statistically significant correlations in large data sets could be considered an enlightening exercise in research methods, and might even lead to some initial observations

that could come in handy at a later stage in a given research project. But as with any collection of empirical data – big or small – one should also recognize the need for social theory to help provide context and guidance in order to separate meaningful relationships between variables from those that are unsubstantial (Silver, 2012).

It follows from this that while the scope of the data – the number of cases gathered and the number of variables employed – is of importance, size is perhaps not all that matters. As suggested by Margetts and Sutcliffe, «Big Data does not necessarily mean interesting data» (Margetts and Sutcliffe, 2013: 141), reminding us not to be blinded by size alone. As such, the quality of the data needs to be taken into account. Perhaps the ways in which data sets derived from social media activity allow for manipulation by the individual researcher should be the focal point. As these data tend to be structured in similar, coherent ways, and as our tools for analysis have grown in sophistication, size becomes an issue primarily with regards to sufficient or insufficient computing power. A large selection of empirical data is of course a good thing, and an absolute necessity in many research settings, but the quality of the data must be considered the first and foremost priority of the individual researcher.

Ethical considerations

Regarding ethical considerations pertaining to this type of research, I will raise three interrelated issues for discussion: (1) the «open» or «closed» nature of data, (2) communicating this type of research to ethics boards, and finally, (3) the need for respondent consent.

First, developments regarding computing power for collecting, storing and analyzing data are not showing signs of stopping or even plateauing. This implies that issues pertaining to the technical limits of the kind of operations that can be performed need to be

discussed in tandem with discussions of which types of activities should be performed We might label this a practical approach to research ethics.

As an example, we can point to some considerations that tend to arise when researching two of the currently most popular social media platforms, Twitter and Facebook. While the services differ in terms of modes of use, privacy settings and so on, we can distinguish between more «open» and more «closed» types of data from both platforms. For Twitter, users can add so-called hashtags – keywords formatted with a number sign (#) that signal a willingness on behalf of the user for their tagged tweet to be seen in a specific thematic context – that can assist researchers as well as other interested users in finding and selecting tweets of relevance. Such uses of hashtags are usually prevalent around the time of specific events, such as political elections, and have served as useful criteria for data collection in a series of studies (e.g. Bruns and Highfield, 2013; Larsson and Moe, 2012, 2013; Moe and Larsson, 2012b). However useful the hashtag criterion might be, there is a need to pose the question of what non-hashtagged content of relevance is available. We can readily assume that tweets that do not include these types of selectors may be of interest to researchers concerned with specific themes. While those types of messages could be gathered by using more open searches, we need to remember the aforementioned issue of the intent of the initial sender. The inclusion of hashtags can indeed be seen as a willingness on behalf of the sender to make the tweeted content publically available within a certain thematic context. While there are other, non-hashtag based modes of researching online political communication (e.g. Ausserhofer and Maireder, 2013), we need to take the open or closed nature of the data into account and shape our research approaches accordingly.

The same reasoning can (taking into account obvious differences regarding the specificities of the platform) be applied when dealing

with Facebook. Arguably a more locked-in service – a user essentially needs to have an account in order to gain access to most of the content – Facebook features Profiles, which is the type of Facebook presence most of us deal with in our everyday uses of the services. While Profiles are mostly associated with non-professional, personal Facebook use, professional employment has recently been taking place on so-called Pages. These Pages differ from Profiles in a number of ways – they are open to peruse by all, including by those who do not have a Facebook account, and they allow their respective owners (in this case, the political actors themselves) to extract more advanced metrics and information regarding usage rates than they would have been able to do if they had employed a personal Profile for professional matters. As with Twitter, we can differentiate between varying degrees of closed or open data here as well, where the operation of a Facebook Page at the hands of a political actor – be it individual parliamentarians, party leaders or even party accounts – could be considered a more open and public approach to the platform, thereby also making our job as researchers interested in the activities of politicians slightly less cumbersome. As general knowledge regarding privacy boundaries on Facebook are generally rather low (boyd and Hargittai, 2010), there might be issues regarding the degree to which activities of individual citizens interacting with politicians in these online spaces should be considered more or less public. The fact that many of the services available for Facebook data collection have built-in, non-revocable anonymizing features for users other than the Page owner should serve at least as a least temporary safeguard against privacy infringements when it comes to research on Facebook Pages. However, as these settings are very much in flux, researchers need to be aware of the specificities of the platforms they are interested in.

Second, the need for research ethics boards has been evident in basically all branches of scholarly activities in order to make sure

scholarly efforts meet the needs and standards set by society at large. While I have dealt with my own experiences regarding the relative difficulty of trying to communicate these issues to ethics boards in a separate, co-authored paper (Moe and Larsson, 2012a), some of these points need to be raised here as well. Essentially, two issues in particular could be raised when discussing ethics boards in combination with Big Data-type research. The first of these concern what could be labeled an «offline bias» in the many forms that need to be filled out when applying for these types of consultancies. As those forms have been constructed to reflect a research environment predominately geared towards research topics far from the specificities of online environments, researchers submitting their applications find themselves having to adapt their own descriptions to offline specifics in order to get the point of the research project across in a correct way. I am not necessarily suggesting that the forms should be extensively rewritten to fit issues pertaining to online research topics exclusively, but rather that those responsible for fashioning these channels for researcher input – be they printed or not – take some time to also adapt them for the many online themes that are currently on the rise within the social sciences and humanities. As such, perhaps these forms could become more specialized for specific types of research. The point here is not that research dealing with the online environment matters differs substantially from offline inquiries; rather, those differences that do exist need to be taken into account when dealing with research projects.

The second issue has to do with the varying degrees of feedback and transparency that characterize the decision-making process of ethics boards. While the information that needs to be submitted to these boards is often plentiful and requires significant amounts of legwork from the individual researcher, the degree to which the submitter gains insight into the reasoning of the ethics board, when they have reached their decision, is of a varying nature. While the

proverbial «burden of proof» should indeed lie on the researcher applying for ethical consultation, we also need to make sure that the feedback received – whatever the decision – is rich enough in detail so that the individual researcher can gain insight into the ways of reasoning applied. By securing at least some degree of transparency in these interactions, and by being more open in communicating these results to the academic community as well as to the general public, we will also be able to move towards precedents that will be very helpful for other, similarly interested researchers.

The third issue has to do with the necessity of obtaining consent when performing research on human subjects. While the practice of securing the willingness of those to be included in your study is more often than not a necessity, the practicalities of performing such operations must be raised for discussion when dealing with certain research projects. As an example, I would like to point to the work performed by myself and colleagues regarding political activity in conjunction with parliamentary elections in Sweden and Norway (Larsson and Moe, 2012, 2013; Moe and Larsson, 2012a, 2012b). While we did employ a hashtag-based mode of data collection, and while this fell in line with the stated regulations (Moe and Larsson, 2012a: 123), the Norwegian ethics board suggested that we attempt to obtain «non-active consent» from the Twitter users studied. We consulted our data sets and quickly realized that such an operation would involve contacting about 9000 Twitter users in order to receive their individual consent. After contacting the ethics board and explaining the situation, we were allowed to move on with our project without shouldering the massive workload of gaining such retroactive consent. Indeed, this signals a willingness on behalf of the entity to enter into dialogue with researchers, arguably a positive starting point. As the data collection was performed only with specifically hashtagged data in mind, this decision could be seen as relatively unproblematic – but we can be sure that there

are other situations – research regarding Internet use by minors, for example – where these issues are perhaps not as clear-cut.

Methodological considerations

As for challenges and questions pertaining to method when performing research on Big Data sets gathered from social media, I would like to raise four points in particular: the problem of «streetlight research,» the access to data, the stability of tools used, and finally, the competencies of researchers.

First, we can broadly conclude that among the many social media platforms available, Twitter and Facebook are (currently, at least) among the most popular, and as such, more interesting for researchers. As suggested by Lotan et al. (2011), the former of these services has indeed become quite popular among researchers – which more likely than not has to do with its accessibility in terms of how it allows for data from the service to be downloaded and archived. As such, the relative openness of the Twitter API (application programming interface) has led to a sizable number of studies dealing with this particular platform. While Twitter has put considerable restrictions regarding API access in place (e.g. Burgess and Bruns, 2012), it still must be considered more accessible than its arguably more popular competitor, Facebook. The relative ease of Twitter data collection, then, leads to what could be described as «streetlight research.» In essence, this is perhaps not a novel problem – we study what we can see, and try to make do with what is made available to us. But if the collective attention of researchers is largely directed at the second most popular service rather than at the one that boasts soaring usage rates by comparison, this could become a problem in the end. Ease of data access might be alluring, but the relative degree to which Facebook has been neglected in the same way creates a knowledge deficiency that does not serve the research community well.

Second, and related to the first point, is the problem of gaining access to data from a financial perspective. As both Twitter and Facebook have started to monetize access to certain parts of their respective APIs, partnering with third-party corporations to handle the day-to-day sales of data, it seems clear that finances will play an ever-increasing role in this type of research. For Twitter, this state of affairs can be illustrated by considering the different types of access allowed. While the so-called «firehose» API – including all the tweets sent through the service – is available, it carries with it a price tag that most academic institutions will not be able to pay. Instead, most researchers make do with what is labeled the «gardenhose» API – which provides a limited stream of tweets for free (e.g. Lewis, et al., 2013). The exact limitations of the latter API have been difficult to ascertain, but it has been suggested that the garden hose variety of access provides about one per cent of the total flow of tweets at any given time (e.g. Morstatter, Pfeffer, Liu, and Carley, 2013). As such, while we should not trust such «gardenhose» access to provide us with all data in a more general sense, more specified searches aimed at limited themes – such as the hashtag-based approaches discussed previously – should provide a fuller, more detailed data set. Of course, this is very much related to the expected scope of the hashtag examined. For Swedish or Norwegian contexts, where few citizens maintain an active Twitter account and fewer still take part in discussing political elections using the service, we can be sure to get a fuller picture than if we were to query the API for data on, say, a hashtag indicating tweets regarding a US presidential election. The key issue here is to be aware of the limitations that the tools employed for data collection carry with them and to shape one's studies accordingly. While a «garden hose» approach to data collection might be able to capture all tweets specified by a comparably limiting selection criterion, it will not be able to compete with the available commercial

services when it comes to the collection of tweets dealing with larger events – such as US presidential elections. Moreover, while such commercial services might be able to provide comprehensive sets of data even for more wide-reaching search queries, the ways in which such data are provided are often not conducive to further research efforts. While these data can provide useful initial insights, a researcher usually wants access to the raw data set. Commercially inclined customers might have these types of «ready-made» analyses as their primary goal, while this is often not the goal of the researcher. As such, even if researchers could afford to access firehose type data, the way they are presented is often not conducive to research purposes.

Third, the stability of the tools we use for data collection and analysis is of the utmost importance. While this is less of a problem for the latter of these activities, where open-source software such as Gawk (Bruns, 2011), Gephi (Bastian, Heymann, and Jacomy, 2009) or Netvizz (Rieder, 2013) sustain large user- and support communities, the former pursuit is arguably a bigger problem. In essence, as both Facebook and Twitter grow ever more popular, the way in which access to data through their respective APIs is granted is subject to more or less constant changes. Take Twitter as an example. As pointed out by Burgess and Bruns (2012), Twitter's changing business model has led to the collapse of a number of free services that were previously available to interested researcher. Facilities like 140kit and the web version of TwapperKeeper were essentially forced to remove certain functionalities from their services, effectively rendering them largely unsuitable for further use. While this problem has been partially solved by the launch of the user-hosted YourTwapperKeeper service (TwapperKeeper, 2010), stability still remains an issue as Twitter keeps changing their modus operandi on a more or less regular basis. If access to data remains unpredictable, this will continue to be a problem for researchers.

As a result of this lack of stability, a number of research teams have taken it upon themselves to build their own tools for data collection. While such tools are often impressive and attuned to the needs of the specific team, the uses of these types of «homebrew» software could lead to what is often referred to as a «silo problem» down the line. If each research team makes use of their own individually constructed data collection tools, ensuring comparability between research results could become a challenge. Although scholars have different needs with regard to the type of data they work with, there is a need for a point of comparison between teams. While total homogeneity is definitely not an ideal, a complete lack of comparative possibilities is definitely problematic.

Finally, the competencies of social scientists and humanities scholars for dealing with these sometimes novel issues of data collection and analysis need to be assessed. Indeed, the need for interdisciplinary efforts is perhaps more pressing than ever before (e.g. Lazer, et al., 2009). If we disregard the aforementioned suggested «end of theory,» the next step is perhaps to further realize what scholars working within the broader confounds of computer science can bring to the table. Theory is needed to pave the way, to find paths through vast quantities of data – but more technical skills are similarly needed to provide access to and manipulate the data in ways that make them receptive to the approaches of the social sciences and the humanities. Such cooperative efforts might not always be easy to carry out (e.g. Burgess and Bruns, 2012), but there is a clear need for them. Perhaps the suggestion by Margetts and Sutcliffe (2013) to provide a sort of «dating service» for different types of researchers could be one way to help bring about these sorely needed interdisciplinary activities. Such opportunities for researchers from different disciplines could be organized in conjunction with academic conferences or other similarly suitable meeting places.

In closing

The move to an online environment for social science and humanities research has indeed been fruitful for both branches. While the above considerations need to be taken into account when planning and executing a «Big Data» research project, they do not amount to a complete list. For example, the «siren-song of abundant data» (Karpf, 2012: 648) concerns the risk of the false impression of representativeness. While gauging Twitter streams and the like for the purposes of public sentiment analysis could prove an interesting methodological exercise (e.g. Groshek and Al-Rawi, 2013), researchers should be wary of making any brash conclusions concerning things to come based on social media data only. As the users and uses of social media in general, and Twitter in particular, tend to be characterized by varying sociodemographics (Hargittai and Litt, 2011; 2012), we must take such sources of bias – often geared towards societal elites – into account.

The issue of stability, mentioned above, is relevant to our understanding of the rather short history of social media platforms. Indeed, while Twitter and Facebook are currently among the more popular social media, this status is destined to come to an end when some novel service is launched and makes its claim for the online audience. With this in mind, researchers need to make sure that their instruments for inquiry – the way questions are posed or coding sheets are constructed – are «stress tested» and stable for future online platforms as well. This is almost certainly easier said than done. As rapid online developments take place, suitably aligned research instruments will enhance the quality not only of our present scholarly inquiries, but also of those to come in the future. Being prepared for these developments might help us in securing longitudinal insights regarding the uses of social media.

This chapter has outlined some of the considerations and challenges faced by researchers studying social media. While my specific starting point has been experience gained from my own

research into online political communication, it is my hope that the topics dealt with here also resonate with those interested in other areas. Finally, it must be mentioned that what has been presented here should not be considered an exhaustive list of issues to be dealt with – ideally, this piece will also serve as a conversation starter for moving on to those further issues.

References

Anderson, C. (2008, 23 June 2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory. Accessed 16 March, 2014.

Ausserhofer, J., & Maireder, A. (2013). National Politics on Twitter. *Information, Communication & Society, 16*(3), 291–314.

Bastian, M., Heymann, S., & Jacomy, M. (2009, May 17 – 20). *Gephi: An open source software for exploring and manipulating networks*. Paper presented at the Third International ICWSM Conference, San Jose, California.

boyd, d., & Hargittai, E. (2010). Facebook privacy settings: Who cares? *First Monday; Volume 15, Number 8 – 2 August 2010*.

Bruns, A. (2011). How Long Is a Tweet? Mapping Dynamic Conversation Networks Ontwitterusing Gawk and Gephi. *Information, Communication & Society, 15*(9), 1323–1351.

Bruns, A., & Highfield, T. (2013). Political Networks Ontwitter. *Information, Communication & Society, 16*(5), 667–691.

Burgess, J., & Bruns, A. (2012). Twitter Archives and the Challenges of «Big Social Data» for Media and Communication Research. *M/C Journal, 15*(5).

González-Bailón, S. (2013). Social Science in the Era of Big Data. *Policy & Internet, 5*(2), 147–160.

Groshek, J., & Al-Rawi, A. (2013). Public Sentiment and Critical Framing in Social Media Content During the 2012 U.S. Presidential Campaign. *Social Science Computer Review, 31*(5), 563–576.

Hargittai, E., & Litt, E. (2011). The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults. *New Media & Society, 13*(5), 824–842.

Hargittai, E., & Litt, E. (2012). Becoming a Tweep. *Information, Communication & Society, 15*(5), 680–702.

Karpf, D. (2012). Social Science Research Methods in Internet Time. *Information, Communication & Society, 15*(5), 639–661.

Larsson, A.O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society, 14*(5), 729–747.

Larsson, A.O., & Moe, H. (2013). Twitter in Politics and Elections – Insights from Scandinavia. In A. Bruns, J. Burgess, K. Weller, C. Puschmann & M. Mahrt (Eds.), *Twitter and Society*. New York: Peter Lang.

Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., et al. (2009). Life in the network: the coming age of computational social science. *Science, 323*(5915), 721–731

Lewis, S.C., Zamith, R., & Hermida, A. (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media, 57*(1), 34–52.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., & boyd, d. (2011). The Revolutions were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions. *International Journal of Communication, 5*, 1375–1405.

Margetts, H., & Sutcliffe, D. (2013). Addressing the Policy Challenges and Opportunities of «Big Data.» *Policy & Internet, 5*(2), 139–146.

Moe, H., & Larsson, A.O. (2012a). Methodological and Ethical Challenges Associated with Large-scale Analyses of Online Political Communication. *Nordicom Review, 33*(1), 117–124.

Moe, H., & Larsson, A.O. (2012b). Twitterbruk under valgkampen 2011. *Norsk Medietidsskrift, 19*(2), 151–162.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K.M. (2013, 2–4 June). *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose*. Paper presented at the 8th International AAAI Conference on Weblogs and Social Media (ICWSM), Ann Arbor, MI, from http://arxiv.org/abs/1306.5204. Accessed 17 March, 2014.

O'Reilly, T. (2005). What is Web 2.0? Design Patterns and Business Models for the Next Generation of Software, from http://www.oreillynet.com/lpt/a/6228. Accessed 17 March, 2014.

Rieder, B. (2013, May 2–4). *Studying Facebook via Data Extraction: The Netvizz Application*. Paper presented at the WebSci'13 Conference, Paris, France, from http://rieder.polsys.net/files/rieder_websci.pdf. Accessed 17 March, 2014.

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail–But Some Don't*. New York, NY: Penguin Press.

Small, T.A. (2012). e-Government in the Age of Social Media: An Analysis of the Canadian Government's Use of Twitter. *Policy & Internet, 4*(3–4), 91–111.

TwapperKeeper. (2010). Your TwapperKeeper – Archive your own Tweets, from http://your.twapperkeeper.com/. Accessed 17 March, 2014.