VERSITA

10.2478/nor-2013-0008
Nordicom Review 33 (2012) 1, pp. 117-124

Debate

Methodological and Ethical Challenges Associated with Large-scale Analyses of Online Political Communication

Hallvard Moe & Anders Olof Larsson

Emerging online arenas offer new possibilities for the study of online communication aided by computer-assisted methods of data collection. However, these possibilities also entail certain challenges. As online data collection such as "scraping" of web content becomes part of the methodological repertoire of non-technically inclined researchers, and as the data available for researchers to place under scrutiny grows ever more plentiful, we point to two challenges that need to be tackled if we are to grasp current developments. How can we make sense of the massive amounts of novel forms of mediated communication, and how can we approach them in an ethically sound manner?

While all branches of media and communication research face these challenges, we are specifically interested in discussing them in relation to the broad field of political communication. Political communication research tries to understand and explain all forms of "purposeful communication about politics" (McNair 2003: 23). To do so, researchers have sampled newspaper articles, studied a selection of television broadcasts, analysed documentaries, or listened in on specific radio programmes. In the age of the mass media, these were perfectly legitimate approaches – and they still are.

Yet political communication research has had to expand its arsenal of approaches. During the roughly 15 years that have passed since the Web emerged as a mainstream platform for mediated communication, researchers have tried to grasp its effect on political communication. The majority of such work, however, notably relies on traditional methodological approaches. If we wish to properly understand online political communication, we need to explore novel possibilities.

To assess the importance of blog posts, twitter messages and other so-called social media outlets as platforms for political expression, to take one pertinent example, we first need to get an overview of the different types of these mediated forms of expression in our polity. This does not only entail intricate definitional delimitations, but also a somewhat daunting methodological task. Those trying to get an overview of the online communication during the latest Swedish election, for instance, would find a lot of material – too much even to get a perspective on, perhaps. The number of messages or utterances users of one not very widespread service (Twitter) themselves labelled #val2010" ("election2010") amounted to over 100,000 during a one-month period leading up to the election. Still, empirical endeavours based on such data constitute necessary steps on the way forward for political communication research. And they clearly entail some challenges.

In the present debate article, we identify and critically engage with two overarching sets of such challenges. First, the scale of the data available for collection and analysis challenge our methodological frames as we collect, sort and study large-scale quantitative data sets – often with the use of computer software. Researchers not only need to learn the practices of new tools for data gathering and analysis, but they must also be able to critically assess the positive aspects as well as the drawbacks of these new approaches. We focus specifically on one aspect of these challenges, concerning data gathering.

A second set of challenges concerns ethics, as the nature of the data at hand often requires us to renegotiate and reflect upon the borders between the private and the public. We aim to identify novel approaches to the study of online political communication, but also to learn from previous research efforts to revise our traditional guidelines and to accommodate these new possibilities in a sound ethical manner.

To illustrate what is at stake, we use an ongoing empirical political communication project: A study into the use of Twitter as a service for public debate in the Scandinavian countries (Larsson and Moe in press). Combining social network analysis of large-scale data sets with more qualitative approaches to study user behaviour, the project seeks to grasp the value of tweeting in everyday political talk, as well as in organized political campaigning during recent elections. Specifically, the project is mainly based on quantitative data on Twitter activity during the national election campaigns in Sweden (2010), and the regional and local campaigns in Norway (2011). Future work will include the national election campaign in Denmark (2011), allowing for comparisons across Scandinavia.

Methodological Challenges Associated with Data Gathering

As the Internet has grown into a ubiquitous part of everyday life in the more affluent parts of the world, and also penetrated parts of the third world, it offers a platform for a hitherto unseen volume of communication and information. This is very much the case for communication about politics. Yet there is not necessarily more political communication occurring now than there used to be. One way to understand the shift is to say that all the informal, ephemeral political communication that used to take place in specific places, at specific times, often in private, and that was never documented – the local activists' street rallies, between colegaues and friends, the discussion around the family dinner table – has all of a sudden become available as posts or saved messages in various online fora.

Volume, then, is an obvious key feature of political communication on the Internet. To understand the functions and values of this immense mass is an equally immense, and also pressing, challenge for political communication researchers. However, as a platform, the Internet does not only facilitate a great deal of communication. it also offers a platform for new forms of data collection.

From the age of mainframe computers to today's sleek laptop varieties, researchers have always made use of information technology to further their endeavours. Along with its connected technologies, the Internet aids historically unprecedented possibilities for searching and retrieving the traces of communication processes. Herein lies an interesting opportunity for researchers, as we are able, with the use of computer software, to collect, systematize and present large amounts of data in ways that did not exist a decade ago.

With these possibilities, however, come some methodological challenges that merit our attention. We concentrate on one such specific challenge, which concerns data gathering.

Among the various so-called microblogging platforms available, Twitter, since its launch in 2006, has grown to become the most popular. By sending short messages – tweets – Twitter users share updates to a network of followers. Tweets can also include so-called hashtags, where the # character is used in conjunction with a word or phrase to connect the tweet to a particular theme. This use of the # sign allows users to search for specific topics of interest, and to follow threads of discussion. Compared to similar services like Facebook, the act of following another Twitter user is not automatically reciprocal. A user can follow any number of other users, although the user being followed does not necessarily have to follow back. As the growth of Twitter use in both everyday and professional situations has become apparent, researchers have taken an interest in the burgeoning service. The immediate problem for these researchers – much like with other forms of online inquiry – has been one of data collection rationale. While there is an abundance of material emanating from the platform (by June 2011, Twitter users wrote 200 million tweets daily (Schonfeld 2011)), the gathering of tweets is not as straightforward as it first might seem. Clearly working in an emerging area of research, scholars interested in Twitter use have approached the matter of data collection in a variety of ways. Some have chosen to download messages directly from the Twitter Application Programming Interface (e.g., Longueville et al. 2009), while others have made use of more user-friendly interfaces to archive the messages of relevance. The syndicated online service TwapperKeeper and its self-installable, free-of-charge variety YourTwapperKeeper, are two such services. As shown by recent projects (e.g., Bruns and Burgess 2011; Larsson and Moe in press; Larsson and Ågerfalk 2011), TwapperKeeper has proven useful for data collection in a variety of different contexts.

These services are free, publicly available online tools that allow for downloading and archiving of tweets according to a variety of criteria. Using a mix of the Twitter Streaming Application Programming Interface (API) and Twitter Search API, the TwapperKeeper tools aim to retrieve every tweet that the researcher wishes based on Twitter user names, keywords in the tweets or hashtags. Specifically, the tools produce downloadable Comma Separated Values (CSV) files, consisting of extensive lists featuring various information and meta-data regarding the archived tweets: the message text, user name and id of the sender, user id of the recipient (if message is a reply), language code, software client used to send tweet, geographical code, and the time the tweet was created. In aggregate, this makes up a rich data set, suitable for statistical analysis as well as different forms of social network analysis. In addition, explorations of such data sets are also valuable when embarking on qualitative analysis. One advantage of this form of data collection is its unobtrusiveness. This does not mean, however, it is shielded from biases. While the approach ensures a comprehensive basis for analysis, it has its problematic aspects.

One first aspect has to do with what is not collected. For data collection connected to a thematic political event or issue, the use of hashtag archives is the best possibility TwapperKeeper offers. In the study of Twitter use during the Swedish 2010 election campaign, we opted to use YourTwapperKeeper to retrieve every tweet, with metadata, that included the hashtag "#val2010". While this resulted in well over 100,000 messages sent by nearly 9000 individuals, this data set clearly did not cover all communication on Twitter about the election. For one thing, given our specific search settings based on the #val2010 hashtag, the tool missed any message pertaining to the election that did not feature the hashtag. Presumably, this means that especially tweets from casual or inexperienced users can be left out. In addition, short replies also tend to not include hastags. As a result, our data set based on hashtag archives risks including a bias towards experienced users.

Moreover, the software itself causes problems. Tools such as TwapperKeeper can be unstable and in need of thorough monitoring. Researchers have found that TwapperKeeper misses out on batches of Twitter activity for limited periods of time, thereby creating holes in data sets (Bruns 2010). Moreover, the researcher needs to take measures to prevent technical malfunctions in the set up of the tool. In a sense, this challenge is somewhat akin to remembering to press record on the VCR and hoping the tape does not jam when archiving political television talk shows. The difference is that no official archive of Twitter activity exist, as it does for television broadcasts. If the software fails, the data from that period are inevitably gone (for further discussions of archival web research, see Brügger (ed.) 2010).

What is often described as the "silo problem" is the second problematic aspect touched upon here. The use of computer-assisted data collection in studies of Internet communication has often relied on self-written software (Park & Thelwall 2006: 8), which creates "silos" of isolated data sets based on non-compatible set ups, and more often than not unsuitable for sharing. This lack of common, available tools that facilitate comparison and enable verification of prior research has constrained the field (e.g., Gaffney 2010). Publicly available, free tools such as TwapperKeeper represent a step forward. A recent move from Twitter, however, has led to a serious setback in this regard.

In March 2011, Twitter chose to employ a stricter enforcement of certain parts of its API terms of service, a set of rules constituting the "codes of conduct" for third-party developers interested in the platform. Specifically, the clauses concerning syndication are now enforced more strongly (Twitter 2011). While the TwapperKeeper service allows anyone to amass tweets based on a variety of different search criteria (the most common one probably being hashtags), the archives created have also been available for download by anyone who might happen upon them. As of March 2011, Twitter has, in essence, forbidden this practice of syndication. Such developments have undoubtedly brought with them problems for interested researchers. At the same time, the YourTwapperKeeper application is hosted and administrated only by a set of specified people, thus effectively circumventing the syndication specifications set up by Twitter. Still, the effect is more isolated "silos", and an obstacle to the development of common methodologies and more stable data gathering tools.

The decision by Twitter to shut down the syndication possibilities led to a series of comments on the blog kept by the team behind TwapperKeeper. A post on the blog regarding the newly enforced restrictions was met with several distraught reactions from users of the service (O'Brien 2011). Thus, the service can be classified as unstable, not necessarily in itself, but because it depends on the rules and regulations stated by the team behind Twitter – rules and regulations that are potentially in disagreement with the functionalities of the service.

It follows from this example that while a service like TwapperKeeper undoubtedly is of great use to researchers interested in the Twitter platform, the relative ambiguity of the rules pertaining to it makes the stability of data collection processes an issue. While the self-hosted open-source variety YourTwapperKeeper is readily available for download and use (TwapperKeeper 2010), we cannot know for certain how long that particular service, or others like it, will be omitted from rules regarding syndication or other, similar areas. To help prevent researchers from tackling this situation by constructing their own interfaces for data collection – thus essentially re-creating the "silo problem" – we should strive to work together towards constructing a common approach – possibly even a common, specific system to be used – in order to provide secure, comparable data collection for future research projects.

Such advice testifies to the budding status of research on online political communication. If we turn our attention from methodological to ethical issues, the impression of a research object in its infancy, and a difficult terrain for researchers to manoeuvre, is further strengthened.

Ethical Challenges: Balancing Public and Private

One way to describe the social sciences is to contrast them to the natural sciences. According to critics of the positivist idea of a common science ideal, social scientists study meaningful phenomena – social action, texts, symbols, or institutions. In Wilhelm Dilthey's famous formulation, this is put down as the division between the natural sciences' aim to explain in terms of cause and effect, and the "human sciences'" aim to understand meaningful phenomena (in Grimen 2003: 65). Max Weber defined sociology as "a science concerning itself with the interpretative understanding of social action and thereby with the causal explanation of its course and consequences" (Weber [1921] 1968: 4). Either way, we come across humans during the course of our work. When we do, we inevitably face ethical questions.

A key interest in political communication research has to do with people's political preferences, information that clearly belongs to the private sphere. As such, ethical awareness has to be an integrated part of our endeavours. In a recent discussion of research on social networks sites, Michael Zimmer (2010) argues that we need to apply a broad, dignity-based theory of privacy (Bloustein 1964). According to such a view,

> merely having one's personal information stripped from the intended sphere of the social networking profile, and amassed into a database for external review becomes an affront to the subjects' human dignity and their ability to control the flow of their personal information (Zimmer 2010: 321).

What is needed when we research use of so-called social media, argues Zimmer (2010: 323), is an understanding of "the contextual nature of privacy" in different instances linked to different services and fora. Users' ideas about what is restricted and meant for the private sphere, and what is in the open and intended for the public, vary substantially, and might not match the researchers' impression, or the service provider's intentions. Much of the discussion of such issues, including much of Zimmer's contributions, has focused on Facebook – a service with notoriously complicated privacy settings, and which invites the amalgamation of different modes of communication that combine private with semi-public and public. Other services constitute other contextual settings.

What if information has been given in a by default public mode of communication by the users' themselves, and identified by the users themselves as political expressions with a thematic label that makes the information searchable? This is the case with the data in our study of Twitter use during election campaigns: Users write a message, include a hashtag that marks the message as relevant to the election, and publish it via Twitter. Using the YourTwapperKeeper service, we wished to collect and archive all tweets tagged so as to indicate political-electoral content during a one-month period before the election dates in Sweden (during 2010) and Norway (during 2011). We were particularly interested in the most active users, specifically who these users were and what particular party or political persuasion they might subscribe to. One might simply argue that the tweets about the election campaign are like letters to the editor in local newspapers. For the majority of users, however, the comparison might not work that well: For one thing, the content, form and style differ. And the sheer number of participants in the Twitter communication also makes the comparison somewhat awkward. More fundamentally, the act of tweeting is different from sending a letter to the editor in terms of its "publicness". To paraphrase Clay Shirky (2008; see also boyd and Crawford 2011), everything that gets tweeted is public, but all of it is not necessarily for the public. Still, we would argue that the setting of our project – thematically tagged communication about an upcoming election – is public, and that the users could be expected to share that view. However, the responses of the institutional review boards involved in our project can illustrate how even this is not a clear-cut case, and also point to some further challenges.

According to the Swedish law regulating research ethics, our scenario of political communication on Twitter was similar to using a questionnaire, for instance. This meant that the institutional review boards' procedure for offline data collection from informants applied to our project. The form to be filled out was notably mostly adapted to various forms of research dealing with offline contexts. The institutions that we as researchers are dependent on need to be alerted to the fact that social scientists and their research projects are moving online to larger degrees – with all the challenges that this move entails.

Our application to the regional ethics board in Uppsala proved successful, and we were given clearance to proceed with our project. This left us none the wiser, however, since as per standard procedure of the board, no specification regarding the reasoning behind the approval was given. This clearly does not make it easier for individual researchers to gain an insight into ethical matters regarding data aggregation. Moreover, when specifications are withheld in this way, the situation and the decision can almost be described as arbitrary. Perhaps more importantly, proper documentation is needed if this case is to serve as a precedent for other, similar research projects. The review board consists of authorities within the field of ethics, but as long as their considerations remain hidden, they contribute little in the way of helping researchers manoeuvre into new territories. Would a similar project aimed at Facebook, for instance, require prior consent or special steps to secure anonymity?

For the data collection in Norway, an application was sent to the Norwegian Social Science Data Services (NSD). NSD acts as a commission for privacy protection for research at all major Norwegian research institutions. Like in Sweden, political preferences are regarded as sensitive information in Norway. Unlike in Sweden, however, NSD had some initial concerns regarding the layout of the data collection. NSD assessed the information itself – Twitter users' messages tagged to thematically link with the election campaign – as "public acts of expression given in an unsolicited manner". This category constitutes an exception to the general rules when studying people's political preferences, i.e. it lifts the major ethically grounded restrictions in relation to data collection and data treatment. Despite this, NSD suggested we "obtain non-active consent". In practice, this would have to be done by contacting every Twitter user who tweeted with the hashtag in question during the period of analysis, informing them about the project, and asking each and everyone of them to get in touch if they opposed their contribution to the study. This would clearly mean a massive workload for the research team: In the Swedish data collection for the project, the body of messages in the data set was generated by nearly 9000 individual users. In contrast to the data collection, which is automated, the task of identifying, contacting and following up these users would have to be done manually. Beyond the workload, such a requirement would obviously entail problems for the comparative potential of the project, as the Norwegian data set, but not the Swedish, would lack those who actively denied participation. Such practical problems notwithstanding, the advice of the NSD clearly illustrates uncertainty: Fundamentally, the data are defined as public, but still – perhaps to be on the safe side – we should seek consent.

In its final decision, NSD withdrew its suggestion regarding consent – in effect reaching the same conclusion as in Sweden, and allowing the project to carry out data collection and treatment as planned. What these experiences tell us is that the ethical issues surrounding research on so-called social media are far from settled. When even the research ethics authorities show inconsistent practices, the challenges for researchers are substantial.

While research ethics for "offline" research could be expected to be quite harmonized, there is still plenty of bureaucratic "red tape" to cut through for scholars interested in online research. As Zimmer (2010; also Boyd and Crawford 2011) argues, we clearly need to educate our institutional review boards about the nature of our research. We call for an international or at least an inter-Nordic harmonization of research ethics law in general, but in particular when dealing with the Internet and with emerging means of data gathering and analysis. In sketching out the details of such a document, policymakers and researchers might find a useful starting point in the ethics guide provided by the Association of Internet Researchers (Charles Ess and the AoIR ethics working group 2002)

Conclusion

Our starting point was a set of challenges stemming from the need to use novel approaches when embarking on studies of online political communication. Illustrated by insights from our own experience with developing a comparative project on the use of Twitter during election campaigns in Scandinavia, we have highlighted two sets of challenges – methodological and ethical – and showed dilemmas that need to be addressed. Our aim has been to raise questions. Hopefully, these questions can stimulate a debate among media and communication scholars in general, who not only need to learn the practices of new tools for data gathering and analysis, but also to be able to critically assess the positive aspects as well as drawbacks of these new approaches. Only then can we contribute to discussions aimed at promoting sound ethical practice.

References

Bloustein, E. (1964) ‘Privacy as an aspect of human dignity: An answer to Dean Prosser’, *New York University Law Review,* 39: 962-1007.

boyd, D. and K. Crawford (2011) ‘Six Provocations for Big Data’. Paper presented at Oxford Internet Institute’s *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, September 21.

Brügger, N. (ed.) (2010) *Web History*. New York: Peter Lang

Bruns, A. (2010) ‘Creating Twitter Timelines from Twapperkeeper Data’. Retrieved from http://www.mappingonlinepublics.net/2010/07/22/creating-twitter-timelines-from-twapperkeeper-data/

Bruns A. & Burgess, J. (2011) ‘#ausvotes: How Twitter Covered the 2010 Australian Federal Election’, Unpublished paper.

Ess, C. & the AoIR ethics working group (2002) *Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee*. www.aoir.org/reports/ethics.pdf

Gaffney, D. (2010) ‘#iranElection: Quantifying online activism’. *Web Science Conf. 2010*.

Grimen, H. (2003) *Samfunnsvitenskapelige tenkemåter, 2. utgave* [Modes of thought in the social sciences]. Oslo: Universitetsforlaget.

Larsson, A.O., & Moe, H. (in press). ’Studying Political Microblogging. Twitter users in the 2010 Swedish election campaign’. Accepted for publication in *New Media & Society*.

Larsson, A.O & Ågerfalk, P. (2011) ‘Right on Track? Corporate Twitter use under pressure’. Paper presented at the NFF Conference, Stockholm, August 2011.

Longueville, B.D., Smith, R.S., & Luraschi, G. (2009) ‘”OMG, from here, I can see the flames!”: a use case of mining location based social networks to acquire spatio-temporal data on forest fires’. Paper presented at the International Workshop on Location Based Social Networks, Seattle, Washington.

McNair, B. (2003) *An Introduction to Political Communication*, London: Routledge

O’Brien J. (2011, June 13) ‘Removal of Export and Download / API Capabilities’. Retrieved from http://twapperkeeper.wordpress.com/2011/02/22/removal-of-export-and-download-api-capabilities/

Park, H.W. & Thelwall, M. (2006) ‘Hyperlink analyses of the world wide web: A review’. *Journal of Computer-Mediated Communication, 8*.

Schonfeld, E. (2011) *Twitter Reachers 200 Million Tweets a day, But How Many Come From Bots?* Retrieved from http://techcrunch.com/2011/06/30/twitter-3200-million-tweets/

SFS (2003) Lag (2003:460) ‘om etikprövning av forskning som avser människor’. Retrieved from: http://www.notisum.se/rnp/sls/lag/20030460.htm

Shirky, C. (2008) *Here Comes Everybody. The Power of Organizing without Organizations*. London: Penguin Books.

TwapperKeeper (2010) *Your TwapperKeeper – Archive your own Tweets*. Retrieved from http://your.twapperkeeper.com/

Twitter (2011) *API Terms of Service*. Retrieved from http://dev.twitter.com/pages/api_terms

Weber, M. ([1921] 1968) *Economy and Society – An Outline of Interpretive Sociology, Volume 1*. New York: Bedminster Press.

Zimmer, M. (2010) ‘”But the data is already public’’: on the ethics of research in Facebook.’ *Ethics Inf Technol* 12: 313-325.