**Using large datasets to study online publics**

Donald Matheson

University of Canterbury, Aotearoa New Zealand

paper presented at New Media and Social Transformation: Convergence Interconnection Share

Central China Normal University, Wuhan, China, 11-13 November 2016

contact

Assoc Prof Donald Matheson

Media and Communication Department

University of Canterbury

Aotearoa New Zealand

donald.matheson@canterbury.ac.nz

please do not cite this oral version without contacting me first

**Using large datasets to study online publics**

The English word ‘public’, and its parallel terms in other European traditions such as the German ‘Öffentlichkeit’ or the French ‘publique’, deal with a hugely important idea – how does society share its most important ideas. If you understand the public, you understand much about the private as well, you understand how politics is done, how power is legitimated. It is one of the most important social imaginaries, to use Taylor’s (2004) phrase, a collective understanding that can be used by people in a society to make sense of social existence. That imaginary has always been an awkward object, difficult to define, and the contemporary public is still more difficult as communication modes become more globalised, fragmented and interactive. What do we mean by this term today? This paper explores the potential of large scale data analysis to help bring this troublesome idea into focus.

In the west, classic liberalism has given us the spatial notion of the public as comprised of individuals coming together in public spaces or in the virtual spaces of public communication to discuss matters of common concern. This deliberative democratic ideal was at least imaginable in early industrial patriarchal western society, where the numbers of individuals in any one society were smaller than today and where a few institutions, such as the legislature, universities and senior Church people were able to embody the idea of bringing together the public or embodying the public interest. Later, it became easier to imagine the public as embodied by institutions of a democratic state and the media, who represent it and are answerable to the individual members of the public in various ways. Yet that version has lost some of its authority too, partly because mass media are much less powerful than even 10 years ago – half of people in the west now use Facebook to get at least some of their news; only 40 percent of 18-24-year-olds eligible to vote in my home country, Aotearoa New Zealand, did so in 2014. The ‘infrastructure of political communication’, as Marcinkowski terms it, is either seriously damaged and in need of renewal or has changed so much in recent years that our tools are no longer adequate. I do not think it is the former. There is more public communication than we have ever had, as I will discuss shortly in relation to the enormous social media response to a major earthquake in my home city. I would like to suggest that the problem is partly one of how the public is theorised and analysed. One way that scholars can contribute to addressing current challenges is to provide society with a better critical reflection of the nature of its public communication.

Before I talk about how to contribute empirically to understanding the public, let us first reflect on some of the problems with the term.

The first is of course that the public is imagined, as Benedict Anderson famously noted. It is not possible to know or even talk to all those with whom we share a sense of participating together in a public. As a result, there is a major risk that scholars or authorities may imagine the public in ways that drift apart from how other people do. A huge amount of power at stake in claiming legitimacy for the news or parliament as connected to the public, or in claiming to be able to know public opinion through aggregating people’s private beliefs via polls. The public, because it is imaginary, is always at risk of being colonised in these ways.

Secondly, the dominant theories of the public build their models from the situations in a few large, rich and powerful western nations. Lowe and Nissen point out that these are in fact atypical of most countries and provide a very poor model for analysis or policy elsewhere. For example, the political culture in my home country of only four million people is shaped by very low social distance between people, where political leaders may literally share the same buses with citizens. The public needs, in Barton Scott’s (2015: 358) words, to be ‘provinciailised’, ‘approaching it less as a normative model for modern society than as a culturally peculiar notion caught up with the particular history of the North Atlantic region (i.e. “the West”)’.

Thirdly, we should be cautious about aiming for theories of a clearly definable, stable public. Charles Taylor (2004) assumes that the public sphere is metatopical, that is, that it does not depend for its existence on the topic of today’s debate but will be there in similar form tomorrow. But historical analysis shows something quite different – publics have always been, in Dewey’s terms, inchoate, relatively unformed and at risk of drifting apart. Dewey’s solution was to combine access to rich understanding of the social world and community-level communication where common concerns could be talked through, building the public out of community. Contemporary media provide examples of publics that remain loose coalitions of interests, rising to accommodate needs and disappearing again, as the western Reclaim movement targeting inequality of wealth and resources did. People also negotiate the bounds of the public on an ongoing basis, mixing public and private modes, and so the actors and gatekeepers for public communication are much less certain than could have been imagined earlier, even in some ideal form.

The challenge is to approach the public in both normative and descriptive ways: too much of one and the concept of the public is at risk of being captured or misunderstood. If we let certain authorities say what the public is, then we are disconnected from and misunderstand the complexity of public life, engaging in a kind of elite colonialism. But then again if we simply describe publics as we find them, then we lose the power to hold those acting in public or for the public good accountable to norms. We allow others who claim to know what the public should look like to fill this signifier with their meaning. It seems to me nonsensical, for example, to ask whether Facebook is a public sphere. That is to fail to separate out real existing relations from the theory. It certainly does not operate by past norms of the public but people are clearly doing public communication on Facebook, in their own ways and according to the dynamics of the platform. Our task is to analyse in closer detail what kind of publicness is done by different people in different media and critique that practice using a range of ideas. This is close to Hannah Arendt’s position, that political philosophy should not look for universal norms, but enable the development of people’s capacity to be true to their own experiences – political theory for her begins with listening and then evaluating in interaction with what we hear.

This less idealising perspective means not looking at the public so much as an institution and more as a performance – as people acting out public cultural practices. Heikkilä and Kunelius (2006) note that the idea of the public as a ‘sphere’, as it if was a space, can divert us from thinking about the action that happens there. Seeing the public as action allows us to see the complex meanings of the public for people and to critique the way it is dominated or organised by certain interests, as well as to see the technologies that shape public interaction. We can also see the particular ways these come together in any one context. Understanding the public therefore depends upon understanding its moral and cultural parameters, which are often complex and contradictory. People are often discursively placed in relation to the public through ideas of what is appropriate to the public and what the public should do through media. Seeing the public as action also connects to theories that highlight how publicness contributes to self-fulfilment, as people become recognised by others as having a contriibution to make to their society. In the Graeco-Roman antecedents of this idea, it is where freedom, creativity and participation in politics are located. People become fulfilled in their uniqueness through freely acting and talking in front of their peers in public – again, this is central to Arendt’s thinking but also to Honneth and many virtue ethicists.

Political culture is hard to study. But the huge datasets that digitised social activity produces now allow the performance of publicness by large numbers of people to be studied in detail. The old texts of the public, official and institutiuonal texts such as government statements, textbooks, news media reports and other forms that represent the public, are joined by digital archives of millions of public interactions. These datasets are well suited to test out contemporary theories of the public, as they are rich with data on interactions and a range of metadata such as location and identity.

Some of the research on this activity is ethnographic, looking closely at the meanings, interactions and identities produced in social media as experienced by individuals, for example danah boyd’s work on young people’s uses of social media to negotiate different social worlds – family, peer group and the wider culture. Much of the research on these new sources of data of public discourse is quantitative, because there is really no other way to process wide streams of material so as to identify significant patterns. Some of that research feels to me rather thin – for example, sentiment analysis that can establish whether positive or negative responses predominate. But some of this research is able to make claims about how individuals and groups orient towards public themes and processes and therefore helps us take forward the theorisation of the public above. I will focus here on Twitter, a social media platform that I have studied and which has some parallels with Weibo in China.

Early research made large claims about the potential of social media to step outside existing constraints on quality public debate. In particular, researchers sought to establish that blogging, social media and other interactive practices allowed public debate to become more deliberative, available to a much wider range of people, freed from the problems of the commodified forms of debate to be found in mass media and freed from the power of an elite to control and suppress. Quite quickly, those claims were found to match actual activity only poorly. Most political blogging, for example, draws upon and interacts with professional news media. Still more problematically, most people’s public activities online are not deliberations on public themes but mix the personal and the public, the rational and the emotional in ways that make it hard to pin it down. What is much clearer is that online public debate operates in ways that are not easily accounted for by classic theories of public debate. We are, therefore at the same point as the theoretical debate above, in needing to empirically establish the forms of public debate online. This is different to analysing who talks and what they talk about. As Schudson noted 20 years ago, rich understanding of the mediation of public life requires us to study the forms – the genres, conventions and tactics – that determine what tends to be said or even what is sayable within any particular technology and the discursive practices that make use of it. He wrote: ‘The power of media lies not only (and not even primarily) in its power to declare things to be true, but in its power to provide the forms in which the declarations appear’ (Schudson 1995: 109).

One major finding about the form of social media publicness is that it is quite individualised. Schmidt (2014), for example, argues that discussion on Twitter operates in terms of what he calls ‘personal publics’. This idea provides a useful starting point for empirical research. Schmidt says that when someone posts about something of significance beyond their personal life, they take part in the public in ways that are referenced according to their own selves. The tweet:

Socialist paradise my arse. @WSJ Babies in Venezuela are dying at a higher rate than in Syria <http://on.wsj.com/2ebesTL> (by Aotearoa New Zealand journalist Matt Nippert)

fits Schmidt’s analysis well. The information is, firstly, selected and displayed according to criteria of personal relevance. It is what Nippert is interested in. Secondly, unlike in mass media forms where the audience is unknown and weakly bounded, it is addressed to an audience of network ties made explicit. Nippert is able to know at least some of those who will read his tweet, because he can track who his followers are and because the responses of some of them will appear underneath the tweet as comments. As a frequent Twitter user, he will orient his tweeting to that known audience. Thirdly, the communication is mainly in a conversational mode, operating at the level of personal opinion, personal relationships and, to an extent, the common stock of ideas that arise in conversational talk. Many other scholars have noted a trend towards networked individualism (Wellman), where people are participating in society somewhat more as individuals than as members of a group or community.

Bruns and Burgess’s (2001) research presents a slightly different picture. Their research finds that Twitter users frequently step into a different mode when they need to step beyond their immediate social circle, by using hashtags. Hashtags allow anyone, not just a Twitter user’s followers, to read the tweets, and at times of major events, such as diastrous floods in Queensland in 2011, a few hashtags quickly become established as ways for large numbers of people to communicate. In these situations, the notion of relevance of material drifts towards the individual’s sense of what is relevant to a much wider group than the self or the personal network, and that network becomes relatively unbounded and the modes of communication combine the personal, a community-oriented voice, formal public announcement and much else. It seems clear to me that public discourse on Twitter is not a stable thing, but changes according to the parameters discussed above – the technology will shape how people engage with wider networks, as will the specific situation, as will the conventions of public life in a particular culture and the institutions of public life in a society. It will also differ significantly according to people’s intentions. I find it more useful to think of people’s modes of participation as tactical. According to parameters of a situation, they will orient to the public in different ways and produce diferent forms of public talk. The great benefit of emerging tools of studying large datasets of public discourse is that they allow us to trace the local differences and specificities of public discourse. In the rest of this paper I would like to present evidence from one analysis that illustrates the strengths of that approach.

Like Bruns, I focus on a moment of disaster to study Twitter’s public uses. Ideas of the public become actualised particularly strongly at certain times, when there is a high degree of reflexivity (that is, when people are conscious that they are participating in mediated events), when there is a sufficient scale to the public communication and when communication contributes in recognisable specific ways to collective social action. Such moments include:

* political events, such as voting, protests, large-scale consultations
* disasters and crises, when social institutions are insufficient
* public-building moments, such as debates about national identity.

At these moments of reflexivity, publicness is both brought to the surface, forming a kind of public consciousness, and is therefore able to be rethought a bit. These moments are not typical, but some of the practices of public communication are particularly visible at the moments.

I have sought to combine the quantification of people’s engagement with close analysis of the specific tactics. Public sphere theory suggests that publicness is only partly a matter of topics and audience, but is significantly a matter of how people interact – orienting towards consensus and operating according to conventions of inclusiveness and rationality, for example. Close analysis of language and context are needed to analyse these things. One tool that allows analysis of language at work on a large scale is corpus linguistics. Corpus analysis is usually focused on lexis – individual words – and their combination into relatively fixed phrases and has led to the development of concordancers, computer programmes that allow those patterns to be studied both as frequencies and in the context of the language around them. The basic idea is that I trace the common language items. I use a concordancer to identify which words are most common, most key (that is, which occur more frequently than in normal written English), the most common phrasal units and the ways that certain words and phrases tend to occur. Following Koller and Mautner, I cycle through looking at common items, then how they are used in specific texts, to build up an analysis.

The 2011 Christchurch earthquake provides a good case study to apply this analysis. On 22 February 2011, the city of Christchurch (350,000 people) was hit by a large and shallow earthquake, killing 185 people, leaving the city centre uninhabitable and damaging 100,000 buildings. The event fits the criteria noted above, a large-scale event which public institutions were not sufficient to deal with and where people were therefore consciously orienting to public forms. There was indeed considerable agency shown in local tweeting, linking to websites where people could get accommodation or find out basic information, linking up students into a volunteer group, called the Student Volunteer Army, who helped clean up, all alongside the official versions and to an extent in tension with them. Twitter became understandable as a space where people were getting things done collectively.

I have used a database of nearly a million tweets that either contained hashtags or keywords related to this and other earthquakes in the region in the year before and after. Of those, 420,000 were produced in the two weeks after the major earthquake and 112,000 were explicitly public through their use of the hashtag #eqnz. I used a concordancer called Antconc to find linguistic patterns in this Twitter data that provided evidence of the tactics of public communication that people deployed after the disaster. I traced, firstly, the most common words, secondly, the words that were most distinctive and therefore characterised the data and, thirdly, traced the language that was widely repeated, including both common phrases and individual messages that were retweeted multiple times. I then closely read tweets and Twitter conversations in which those textual elements arose. I did not analyse the many news reports whose headlines were posted on Twitter, as they operate by a particular understanding of the public that was to the side of this project. The strength of this approach is that I was able to study the text in its context of meaning, as part of a flow of tweets that users would have seen if they had followed people tweeting about the quake or looked up the #eqnz and other hashtags.

There are methodological problems with this approach. There is no bounded Twitterverse, and without a reliable network analysis of what public communication different Twitter users had access to, there is some risk that the dataset is a little unreal. No one actually could experience the set of 420,000 tweets I analysed and everyone’s experience of the public talk around the earthquake would have been a little different. But the focus on the forms of discourse rather than any particular content reduced this problem.

The findings here diverge from two central threads of the growing post-disaster Twitter literature, one of which focuses on the reliability of information produced via Twitter and the other on how social media works as a self-organising system that takes over government is weak. Both of these imagine Twitter as a measurable public space and as performing the role of a public institution – as alternative media, or as a disaster response mechanism. My data provided evidence that people did not use it as a public forum as defined in classic western liberal theory. I found very few tweets that spoke on behalf of the city or its people, taking on the kind of authoritative public voices that a leader might, as in the following:

CHCH does not need advice on how to rebuild, it needs time to clear up and wait to see if this fault is permanent or not #eqnz

Rare tweets of this kind were still more rarely retweeted by others, tending to be met by silence. We can conclude that these tweets did not resonate (in Zhou et al 2010’s terms) with local Twitter users’ ideas of public talk on Twitter at this moment. Twitter was not a platform for the clash of individual views. As I will discuss further later, this absence fits a picture of New Zealand political discourse where a little more collective rationality is preferred over individual rationality, and whose modes are more consensual.

There was some evidence of Twitter being used by people to organise horizontally when the hierarchical structures of normal society fail to function in the disaster context. On the day of the earthquake, an older couple who were trapped in a hotel tower (room 2302 of Grand Chancellor) could not get through to the emergency services and instead called their son in India, who sent requests out on Twitter. A New Zealander Rob Thompson called emergency services, then tweeted about it, and his tweet was picked up by various people, including a television reporter who was with rescuers. The rescue authorities then sent someone back into the hotel (Bussmann 2011). That story, while unique, is the kind of story that feeds an understanding of Twitter’s value in which individuals reach out to other individuals across the network when they were not confident of public structures – that is, as a self-organising system. People used Twitter in this way to search for missing people, to alert people about tools to find others or to find accommodation and above all to pass on information. The following is typical of this kind of tweeting:

RT @[user]: Bob Parker: Schools closed for the week, as are workplaces. Stay at your homes, or help in your communities. Don't drive ...

A number of users with larger followings, such as celebrities and technology innovators, took on a role of retweeting public announcements. Nearly half of all tweets in the sample are retweets and a similar proportion contains internet links; a major use of these functionalities of Twitter was to pass on information. Close analysis of the data, however, shows that there was little apparatus of formal public talk or metatalk about Twitter, which might be expected when individuals are usurping public functions that typically belong to institutions. The talk seemed to be understood in a more local sense of people helping each other. This is partly to do with the medium: Twitter’s syntax lacks hierarchy, so that the mayor Bob Parker’s announcements and retweets of them appear in exactly the same way as statements by individual citizens. But the lack of formality is present in many other ways: for example, the missing search talk often uses the informal phrase, ‘Anyone know…?’ The self-organising among Twitter users happened with very little politicisation but was instead pragmatic and informal in its language. The rarity of commentary and metatalk about Twitter as a forum of sharing is consistent, again, with a consensual or collaborative idea of the Twitter public.

The single most dominant form of tweeting comprised expressions of care and offers of help, operating in terms of an abstract and globalised public. In the first two days, a large number of individuals expressed their solidarity with victims – 5 percent of all tweets on the first day included the word ‘thoughts’, often in the formulaic phrase, ‘our thoughts and prayers’. Some of these tweets came from celebrities and sports stars, such as the Formula One driver Jenson Button or the actor Alyssa Milano, and these tweets were often retweeted hundreds of times. These tweets usually used the abstract term, ‘people’, as in ‘the people affected’. This phase of response was quickly joined by a huge wave of formulaic calls to donate funds, which were again widely retweeted, and tweets aimed at celebrities asking them to retweet calls for donations, such as:

RT @[user] Please can you RT? Christchurch NZ needs your help! Please helpquake victims http://bit.ly/hpufB or ...

In these ways, an instant public was widely imagined by Twitter users, that was both globalised and abstract, and which operated on humanitarian, emotional and popular cultural terms. This kind of public discourse was almost universally one-directional: people did not reply to these solidaristic tweets, which makes them statements in the presence of a wide public rather than talk between actual people. Indeed, when individuals who had more personal relationships with the earthquake victims expressed solidarity, they expressed their thoughts and prayers in terms of ‘family’ and ‘friends’ rather than to the abstract term, ‘people’. We could term the norm underpinning the more general call to people a ‘for-everyone structure’, in that it encompasses everyone listening. An explicit example would be:

Tell everyoneto change their voicemail to let people know they are okay, if their battery is running low! #eqnz / Great idea! #chch

This operates in a similar way to how Durham Peters describes broadcasting, as an act of faith in an unheard public. Thus, while it is an act of sharing in public and while it is often quite an individualised act by one person showing her or his care, it is not an interpersonal one but sent out to everyone.

These tweets dwindled very rapidly after the initial news of the disaster. By day three of the Twitter sample, and particularly among those in Aotearoa New Zealand, a common way of talking emerged that I would call negotiated statements. These were tweets where users’ address to others was more complex, not assuming the tweet’s relevance or importance, presenting a less self-assured tweeter, using less formulaic and less formal language and not assuming a universal addressee. They were characterised by phrases such as ‘If anyone’ or ‘If you’, as in:

If you'rein Chchand you have no toilet or portaloo I now have 300 chemical toilets in my garage #eqnz drop me a tweet

It is, I think, noteworthy that the tweet does not simply inform by telling people, ‘I have 300 toilets,’ or call on people to act, ‘Contact me to pick up a portable toilet’, but chooses a conditional structure (‘if...’). This group of tweets often linguistically marked participation in a group of those affected, and thus were insider messages, affiliating to the group, rather than claiming to speak on behalf of the public or addressing them. For example,

Awesome, the combined effort of everyonehas just tipped the NZ Red Cross 2011 Earthquake Appeal over the $1 million mark. #EQNZ :-) ^ME

which was written by a public relations account belonging to a large business, chose phrasing, ‘the combined effort of everyone’, that did not address ‘you’ but instead assumed shared membership, as well as signalling that the group had an informal national identity through the colloquialism, ‘Awesome’. The tweets were widely retweeted within the country. I would characterise them as a ‘for-anyone-as-someone’ structure, drawing on Scannell’s (2000) characterisation of British public broadcasting as neither addressing people as specific individuals (someone) nor as a simple mass (everyone), but recognising the realness and specificity of the individuals listening. In fact, the tweets do that more strongly than broadcasting can, for they recognise the unseen people’s specificity through the tentativeness of the claim to be relevant to them. In Benhabib’s terms, people are acknowledged as ‘concrete others’, while being indefinite still. This kind of address makes sense in a small country, where people are likely to know some of their addressees, but it is also culturally appropriate. In Aotearoa New Zealand, it can be seen as arrogant or bossy for an individual to take on too much public authority or to impose on others, and the claim to knowledge or the emotional intensity of the message is often negotiated down. This kind of language was widely taken up by official voices later, who recognised it as a culturally appropriate way of doing publicness on Twitter.

Some of the tweets that were shared most widely in the post-crisis period referenced shared experiences. These tweets were not so strongly coherent linguistically, other than a few humorous memes which played with variations on a set phrase, ‘You know you’re from Christchurch when...’ They tended to cohere more through their requirement that the reader bring specific local knowledge to bear, including deixis (such as ‘that’ to refer to an earthquake that had happened seconds before), through humour that carried a claim to capture the public mood, or through reference to individuals, such as the city’s mayor, by name without explanation. They acted as a collective self-representation by people in the city. They also acted out a sense of togetherness, according to dominant cultural tropes, including self-deprecating humour, understatement and casualness. A widely observed phenomenon after the disaster was a rapid increase in people’s sense of community, as neighbours helped each other or provided services to others for free (TEPHRA 2012). This was clearly replicated in Twitter, but in doing so it produced a public discourse that tended to favour a particular kind of community. The shared references brought people to the same level by privileging the shared experience and did not differentiate between politicians, celebrities, corporates and citizens. ‘Please RT when you can,’ a New Zealand sportsperson tweeted to the Prime Minister about a fundraising auction of his own sportsgear. These texts were often pleasurable, celebrating survival in a subversive, levelling kind of public togetherness – something that scholars found also in post-disaster Haiti (Wagner).

So what social imaginary of the public emerges in people’s communication in this post-disaster moment? We should discard immediately assumptions that the public was performed primarily by leaders or through publishing factual information or through rational debate. These traditional markers of a mass-mediated public sphere do not match well onto the ways that individuals used Twitter. Nor do they capture what is distinctive about public communication in New Zealand. Instead there were multiple overlapping ways of being public, which tended to be both individualised and oriented towards mutual regard for other individuals, but which differed widely in their global or local focus. This public activity can be quite successfully read as moral publicness. Much of the discourse can be read in Arendt’s terms as moral self-improving – people become better versions of themselves through caring for others and being creative in a public setting. Some of the communication was also governed by local cultural norms of humour, consensus, politeness and low social distance. This is partly to do with the disaster – at a later moment, we would perhaps see more political discussion and less mutual regard. But this kind of analysis also brings to the surface aspects of an enduring, underlying political culture through which people responded to the event. This means also that Twitter did not produce its own public space. Assumptions that Twitter operates as a self-organising system after disaster were only borne out to a small degree and the discourse referenced pragmatic, informal and consensual ways of operating that are valued in the country. The hashtag #eqnz was an important connector but it was only part of the logic of this public. Indeed, rather than read this kind of communication as evidence for one kind of public, as evidence of a collective response or structure, we need to get better at reading large scale datasets as highly fragmented and multiple. The tactics of publicness here were not a collective act, not a coherent set and not specific to Twitter, but partly a matter of actualising ideas of how public talk works that are also performed elsewhere.

Computer-assisted analysis of a large dataset such as this provides insight into the way people act in public ways in contemporary societies. The analysis here was qualitative, although beginning with the larger scale patterning that the concordancing programme helped to detect, and that seems important to me so that we avoid reducing public talk down to its content or to networks or to other aspects that can be fully quantified. For the public exists at the level of people’s imaginations, motivations and interactions in particular situations, qualities which cannot easily be counted. The empirical challenge is to study the digital traces of publicness in a wide range of settings and cultures, and I would suggest that methodologies such as the computer-assisted discourse analysis done here, which allow meaningful forms of public communication to be studied at scale, will be valuable in taking forward the task of reimagining the public.