A 'Big Data' Approach to Mapping the Australian Twittersphere

Axel Bruns, Jean Burgess, and Tim Highfield
ARC Centre of Excellence for Creative Industries and Innovation
Queensland University of Technology
Brisbane, Australia
a.bruns / je.burgess / t.highfield @ qut.edu.au
http://mappingonlinepublics.net/

Introduction: Twitter and 'Big Data'

The widespread adoption of leading social media platforms such as Facebook and Twitter in much of the developed world has also led to a rise in research projects across the humanities and social sciences which seek to investigate and analyse the emerging uses of these platforms. A substantial number of such research projects have applied existing communication and cultural research methodologies to this task: this includes both qualitative approaches (for example, the close reading of textual and communicative artefacts sourced from these platforms, or the ethnographic study of specific users and user communities) and quantitative methods (such as surveys of users to examine their attitudes and activities, in order to explore larger behavioural patterns).

However, the increasing – if not always comprehensive – availability of structured data on user activities through the Application Programming Interfaces (APIs) of such social media platforms has also enabled the development of a range of new research methods which see the intersection between humanities-oriented Internet studies and the 'big data' paradigm and accompanying debates (Burgess & Bruns 2012). Research which builds on such API data is potentially able to access much larger datasets on user activities and to develop more systemic perspectives on the uses of social media than previous approaches have been able to do; alongside other developments in the 'new' digital humanities this is seen by some scholars as representing a "computational turn" (Berry 2011; 2012), while others portend the development of more scientistic models in media, cultural, and communication studies (for example, Manovich 2012; Hartley 2009).

Social media platforms provide API access to user activity data in the first place to enable the development and use of third-party tools and applications which are able to enhance the user experience and provide additional functionality (for example, to high-end and professional as opposed to casual users). However, the availability of such data has also supported the development of social media analytics as a field of industry as well as scholarly research. Such research has been able, *inter alia*, to examine the uses of social media for everyday communication (for example, boyd *et al*. 2009; Papacharissi 2011; Marwick & boyd 2011), as a backchannel to live and televised events (Deller 2011; Dröge *et al*. 2011; Weller *et al*. 2011; Harrington *et al*. 2012; Highfield *et al*. 2013), in crisis communication (Lotan *et al*. 2011; Bruns *et al*. 2012; Mendoza *et al*. 2010; Palen *et al*. 2010;), and for political campaigning (Bruns & Burgess 2011; Burgess & Bruns 2012a; Christensen 2011; Larsson & Moe 2011; Stieglitz & Dang-Xuan 2012).

In the case of Twitter, a substantial amount of existing research which draws on the Twitter API deals with relatively discrete case studies. Such case studies typically capture tweets relating to specific Twitter hashtags (keywords prefixed with the hash symbol '#', which enable users to search for and track topical conversations more easily), and usually draw upon datasets which range from several hundred to a few million tweets. The 'hashtag studies' approach is especially useful for the analysis of key communicative activities in the context of identified events or topics: to include a hashtag in one's tweet constitutes a process of self-selection on behalf of the user, so hashtag datasets will contain only those tweets which their authors felt were relevant to a specific topic, from users who were familiar with the hashtag concept overall and knew the appropriate

hashtag to use for their specific topic. For tweeting about one-off events, especially televised content, and subjects which strongly promote their related hashtags (such as the Occupy activist movement or crisis relief efforts), focusing on hashtags is an appropriate methodological choice, since communication is directed towards these markers.

While these hashtag-based datasets can be very large (far too large for a single researcher or even a small team to deal with using manual content analysis, for example), and a proper treatment of the sociocultural phenomena to which they relate can be very 'thick' indeed, the need to delimit the corpus in advance (through the selection of the hashtag) means that such studies rarely provoke the same epistemological and methodological shifts as more radically data-driven 'big data' studies do. Further, hashtag datasets are unable to capture the wider communicative activities around the topic or event described by the hashtag (for example, posts between individual users who refer to the topic, but chose not to make their tweets visible more widely by hashtagging them); while little comparative research into such matters has been conducted so far, our own explorations show that four times more tweets contained the keyword 'tsunami' than the hashtag '#tsunami' in the immediate aftermath of the major earthquake in northeastern Japan in 2011, for example (an unknown and unknowable, but still larger number of tweets would have referred to the event in yet other ways, without using 'tsunami' or '#tsunami' at all). Even where hashtag datasets like #tsunami or #occupy grow very large in their own right, they continue to represent only a self-selecting subset of all tweets which are related to their areas of interest.

Further, even when a prominent hashtag has developed for recurring topics of interest on Twitter, the longevity of these hashtags means that patterns of use have changed in ways which do not necessarily affect the briefer, more focused discussions in response to televised events, for instance. The #auspol hashtag, denoting tweets about Australian politics, is a prime example of this; while the hashtag remains widely-used, its notoriety for inflammatory comments and replies, for trolling, and for the dominance of a small number of overwhelmingly active users, has led other Twitter users with an interest in politics to not include it in their own tweets.

A hashtag-based research approach also depends crucially on the existence of a common and widely adopted hashtag in the first place, of course. Hashtag research is especially valuable, therefore, in the context of foreseeable events (with foreseeable hashtags), such as elections, sports, conferences, or media programming, or for unforeseen events for which a dominant hashtag quickly emerges (such as some political crises and natural disasters). It is far less able to address more general themes of discussion (music fandom, the state of the economy, the progress of scholarly research) unless a group of lead users in any such area promotes a preferred hashtag for the discussion of their interests – and in doing so, erects a new barrier between committed and more casual users, similar to the #auspol example above. Hashtag datasets are therefore especially unable to shed light on everyday communicative activities which are linked only loosely to any one specific topic – yet such everyday activities constitute a substantial part of Twitter activity and (if researched) may offer valuable new insight into fundamental communication patterns on Twitter.

Thus, perhaps the most significant limitation of hashtag datasets for the study of Twitter and its uses is the fact that these datasets necessarily contain all relevant tweets matching the hashtag or hashtags chosen for analysis (subject only to server failures or API rate-limiting, they are comprehensive), but provide no information whatsoever on the broader communicative context surrounding these hashtags on Twitter. A hashtag dataset on a major event may contain several hundred thousand tweets, but it is impossible to determine from the dataset itself whether at the height of its activity the hashtag was widely visible across the entire Twitter network (for example as a 'Trending Topic' – a hashtag or theme identified by Twitter itself to have experienced a rapid growth in volume), or whether it represented one of several other major themes being discussed on Twitter concurrently. Indeed, beyond the limited and somewhat vague information

occasionally released by Twitter itself – for example, Usain Bolt's 100m final at the 2012 Olympics generated over 80,000 tweets per minute (Twitter 2012a) – it remains very difficult for researchers to place their observations about individual hashtag datasets in a wider context. If the 2010 Australian federal election on 21 Aug. 2010 generated 94,000 tweets on election day alone, for example, does this represent activity by a considerable percentage of all Australian Twitter users? How does election-day tweeting compare with tweeting during the final episode, say, of popular reality TV cooking show *Masterchef*, or the Rugby League or Australian Football League (AFL) grand finals?

One approach to addressing such questions is to pursue a greater number of comparative studies across hashtag datasets; this also requires the development of a range of standardised metrics for the quantitative description of Twitter user activities, to provide a reliable basis for such cross-comparisons. Such metrics (Bruns & Stieglitz 2012; 2013) enable the identification of stable patterns of user activity across comparable events, indicating that communicative patterns are different when Twitter is used as a backchannel for televised events, for example, compared to its use in crisis communication. But even this does not address the more fundamental problem of comparing hashtag activities against the full range of public communication which takes place on Twitter at any one moment.

Most fundamentally, hashtags are unable to shed much light on how everyday users may encounter information on Twitter. While it is possible for users to deliberately track the content of hashtags themselves, of course, and thereby to encounter tweets even from other users whom they had not previously been aware of, this serves as only one possible layer of communicative exchanges on Twitter (Bruns & Moe 2013). By contrast, the most common and most central path of message transmission on Twitter is through the networks of interconnection between individual Twitter users (as followers and followees of one another, and not necessarily in a reciprocal connection). Hashtag datasets provide information on what tweets were marked with a hashtag, but their packaging as a unified dataset is a convenient fiction for the researcher rather than representing the lived experience of most Twitter users, for whom hashtagged and non-hashtagged tweets appear alongside each other in their Twitter feeds. Such datasets cannot tell us how many users such tweets may have reached; information on the shape of follower/followee networks on Twitter provides a more direct understanding of how far individual messages would have been visible throughout the platform.

Additionally, in many cases it will also be desirable to understand user activity across hashtags not against the backdrop of all other Twitter communication, but rather in more local (for example, national contexts). Australian uses of Twitter, for instance, are likely to be dwarfed by the total volume of tweets being posted at any one point in time, and such comparisons are therefore less than useful. It would be more appropriate to establish a baseline of everyday Australian user activity, filtered from the total 'firehose' of global Twitter communication, in order to situate individual hashtags within this national context. Possible in principle, if difficult in practice due to the increasingly commercialised nature of large-scale Twitter data (an issue which we return to in the final section of this article), such a comprehensive identification and analysis of the public communication activities of Australian Twitter users would constitute a 'big data' initiative in its scale, and a more 'data-driven' one in its approach.

Such an approach, in turn, crucially relies on the identification at least of the majority of Australian Twitter users, so that their tweets can be extracted from the global firehose of tweets. This constitutes a substantial methodological challenge in its own right: while profile information about individual users is readily available from the Twitter API, alongside their tweets, there is no direct way of retrieving a list of all Australian users from the API, nor is there even reliable data about the full extent of the Australian Twittersphere – available industry estimates (which rarely identify the rationale for their figures) range between 1.8 and 2.2 million Australian accounts (for example, Bull 2010).

In response to these challenges, this chapter outlines our approach to developing a reliable picture of the Australian Twittersphere, demonstrates the utility of these baseline data in shedding new light onto existing hashtag datasets, and maps further steps towards a more comprehensive analysis of Twitter use in Australia. Finally, we position this work as an example of the computational turn in the digital humanities, and discuss the practical and conceptual challenges which lie ahead.

Mapping the Australian Twittersphere

Studying the wider Australian Twittersphere provides the opportunity to look beyond isolated topical and event-based communication on Twitter, by instead examining how these discussions overlap across a national population of Twitter users. Do users who tweet about politics, or parenting, or sport, primarily follow users with these same topics of interest? Are a user's home town and state, for example, also factors which may be reflected in who they follow on Twitter? A wealth of new information comes out of such research; as the first detailed study of the Australian Twittersphere, the identification of the size of the Australian Twitter userbase, and its topics of interest, invites further exploratory and comparative work around other national Twitterspheres – and their interlinkages. To date, while research into communication on Twitter has covered a variety of contexts, as noted above, attempts at mapping the structural follower/followee relationships between a national Twitter population, in particular identifying communities of interest and comparing these implicit topical clusters with the *ad hoc* publics arising from tweeted discussions, are more limited. Our research into the Australian Twittersphere, then, also establishes a framework which may be adopted for future studies of intra- as well as international Twitter activity.

The first step in developing a more comprehensive picture of the Australian Twittersphere is to identify individual users as Australian, within acceptable margins of certainty. Twitter provides users with a number of means to identify their location or nationality: they may do so in the free text description which is attached to their profile, state the town or city where they are located, or provide geolocation coordinates which are attached to their Twitter profile. These datapoints are of limited use for our purposes, however, given their inconsistent use: descriptions provided by the users themselves may be vague or ambiguous, for example – users may say they are in 'Australia, 'Victoria,' 'Perth,' or 'Paddington,' but several such names could also point to locations outside of Australia; or they state their location as 'Oz,' 'Out West,' or 'BrisVegas,' or a wide range of other alternatives which could not always be reliably identified. Indeed, during the protests against manipulations of the 2009 Iranian election, many Twitter users worldwide set their location to Tehran or other Iranian cities (Burns & Eltham 2009), in order to make it more difficult for Iranian authorities to identify domestic users; free text location information may be gamed, therefore. Geolocation, on the other hand, is not widely used on Twitter (Wilken 2013), and would identify only a small percentage of the total Australian user base.

A further datapoint available through the Twitter API provides a more useful alternative: in setting up their profiles, users are also able to set their home timezone; due to the diversity in Australian timezones and daylight savings regimes, Twitter provides specific timezone options for Perth, Adelaide, Darwin, Brisbane, Canberra, Hobart, Melbourne and Sydney. Perhaps not least also because this setting appears in fourth place in the Twitter Website's 'Basic Account Settings' page (after the username, email address, and language settings), ahead of many more minor settings options, a substantial number of Twitter users do appear to make use of this customization option, as we will see; in turn, this provides a means not only to identify users as Australians (or at least as based in Australia), but even to pinpoint their likely home state.

Our approach to identifying Australian Twitter accounts is premised on the use of the timezone information in each user profile, therefore; this provides a clear and unambiguous datapoint which can be captured and evaluated by automated means. Using this datapoint, we could – in theory – request information about each

and every Twitter user from the Twitter API, and divide Australian from non-Australian accounts. However, this brute-force approach would be highly wasteful, as it would require us to test several hundred million Twitter accounts in order to find an estimated two million Australian accounts. Instead, based on the assumption that most Australian Twitter users are more likely on average to connect with other Australians than with international accounts, we proceed by performing a snowball crawl of the Australian Twitter network; this also has the advantage of generating additional information about the shape of that network, which will be valuable in its own right.

Over the course of 2011 and 2012, we engaged in such a snowball crawl of the network. We began with a seed list of Twitter accounts which participated in hashtag conversations on particularly Australian topics, including #ausvotes (for the 2010 Australian election), #auspol (a continuing discussion of Australian politics), #qldfloods (for the 2011 south-east Queensland floods), and #masterchef (for the popular television cooking program). Accounts appearing in these hashtag datasets were first tested to examine whether they had set an Australian timezone in their profiles; for those accounts which were accepted as 'Australian' by that criterion, we then retrieved their follower and followee lists. (Any 'private' accounts, whose tweets and network connections are not publicly available, were ignored in this process.) We then tested the timezone settings of each of these followers and followees in turn, and repeated the process with each newly identified account which had set an Australian timezone.

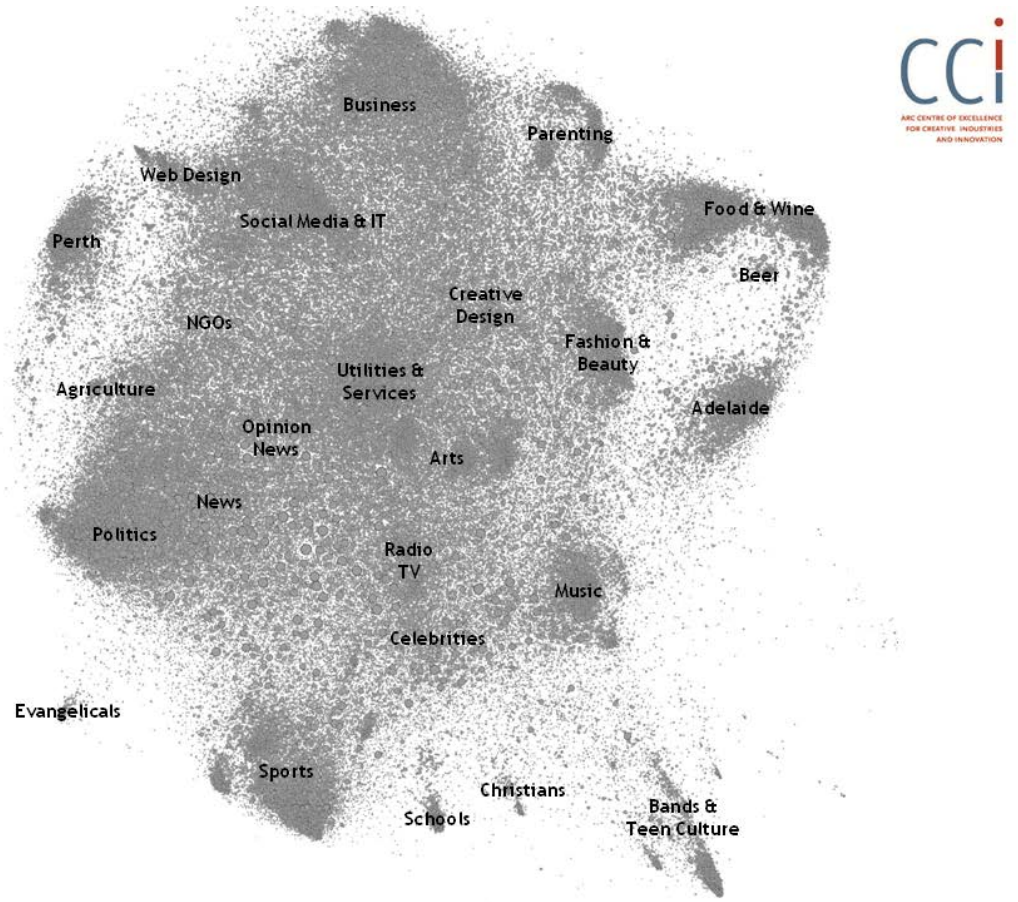
Due to significant access restrictions which Twitter places on its API, this is a slow process: at present, only fifteen follower or followee lists may be retrieved from the API during any one fifteen-minute window (Twitter 2012b), and for users with more than 5000 connections, the API counts such retrieval procedures as multiple requests. (These rate limits have changed several times over the course of our research project.) To date, we have tested more than four million Twitter accounts for our timezone criterion, identified more than one million of these accounts as Australian, and retrieved their follower/followee connections. The network crawl continues at the time of writing, and the decelerating rate at which we are identifying additional Australian accounts is consistent with a total population of some two million Australian Twitter accounts. However, on the basis of the data retrieved so far, identifying both the accounts themselves and their network connections, it is already possible to sketch the overall shape of the Australian Twittersphere as we have constructed it here – with the rather large caveat that there is no necessary correspondence between the topography of this network and any user's lived experience of a Twittersphere, or 'networked public', from a (literally, not pejoratively) egocentric point of view (boyd 2010).

In the first place, based on the network information which we have retrieved, it is possible to visualize the total network of interlinkages between accounts. To simplify the discussion and visualization of results for the purposes of this chapter, we concentrate here on those accounts which have at least twenty incoming or outgoing connections to the rest of the network (i.e., degree = 20 or above). This (necessarily arbitrary) cutoff removes users from the network who have yet to follow or be followed by more than a handful of others, including especially very recently created Twitter accounts whose connections are still forming; the removal of such recent users, in particular, also removes any distortions of the overall network structure which the inclusion of such still-nascent follower relationships would introduce. What remains, instead, are the most connected 120,000 of all identified Australian accounts.

Using the Force Atlas 2 algorithm provided by the network visualization software *Gephi* (Gephi.org 2012), which simulates gravitational attraction between connected nodes in a network, we arrange these accounts in line with their strength of interconnection with other accounts: accounts which belong to a cluster of highly interconnected accounts will be placed close to those clusters, while accounts and clusters which share few common connections will repel one another. Force Atlas 2 is especially useful for our purposes because it is able to visualize large and complex networks, and is optimized to highlight densely interconnected clusters in

the networks it visualizes; this is important for the visualization of the large dataset of Australian Twitter users' connections which we are dealing with here. It should be noted in this context, however, that a large number of other network visualization approaches are available both within *Gephi* and in other network visualization tools, and that there is no one 'true' representation of any network dataset; our results using Force Atlas 2 provide one approach to viewing the network, but other visualizationss are also possible. In Jacomy *et al*. (2011), the authors of the Force Atlas 2 algorithm discuss the respective advantages and drawbacks of their approach in comparison to other popular algorithms. We refer here also to critical work on the performativity of algorithms that construct and represent networks such as that of Tarleton Gillespie (2013) – while Gillespie is writing of algorithms such as that used by Klout to calculate popularity, in relation to network visualisation, too, we must note 'the friction between the "networked publics" forged by users and the "calculated publics" offered by algorithms' (Gillespie 2013 n.p.). In what follows, then, we attempt to represent our network visualisations as tools for exploration, not as literal or total representations of social realities.

Several network clusters and multi-cluster structures emerge through this network analysis process. Further qualitative exploration of these clusters, which examines the accounts that are central to each cluster, is able to provide a rationale for each of these clusters; for example, one densely interlinked section of the network contains the leading Australian news organisations, political journalists, politicians, and political activists, and further subdivides into areas which are dominated by various conservative and progressive party interests, or feature accounts related variously to social policy, environmental concerns, or agricultural interests. Another is dominated by sports-related accounts, and subdivides into a range of different sports and sporting codes. Figure 1 presents this preliminary map of the Australian Twittersphere and labels the different clusters, while Figure 2 provides additional detail by indicating the subdivisions in the large politics cluster.[1]



[1] Higher-resolution versions of all graphs in this chapter can be found at http://mappingonlinepublics.net/2014/01/01/repurposing-appendix/.

Figure 1: Thematic Clusters in the Australian Twittersphere. Based on network information for the 120,000 most connected users in the overall network.

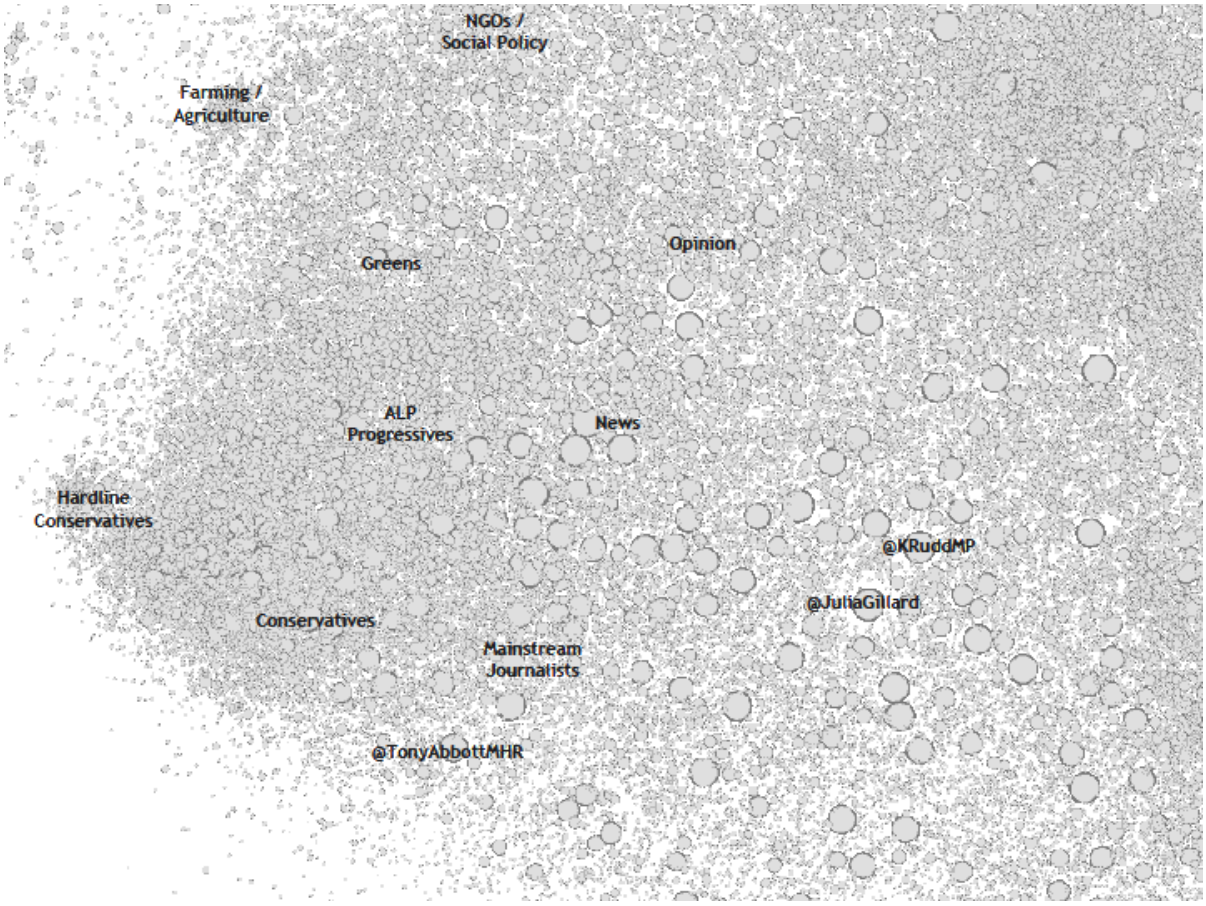


Figure 2: Subdivisions within the politics cluster of the overall network map.

In discussing the map in Figure 1, it is important to note again that it is based on follower/followee network information, not on data about these accounts' tweeting activities. It indicates that clustering in follower relationships in the Twitter network is based largely on thematic affinity, rather than on family relationships or personal friendships (as may be the case for a network such as Facebook), and provides a reasonable approximation of the major themes of everyday Australian Twitter discussions, and their relative centrality to the network: politics, business, lifestyle, the arts, and sports, subdivided into a range of smaller and more specific themes. This does not mean that the accounts placed in each cluster *exclusively* discuss such themes in their day-to-day activities, nor that they *only* link to other members of the same cluster; nor is there any reason that users who do not belong to a specific cluster would not occasionally touch on themes which that cluster would cover more frequently. The small 'beer' cluster in the top right of the map, for example, represents a tightly interlinked network of craft breweries, specialist retailers, and related accounts whose primary *raison d'être* for networking is beer culture; other accounts in the network may also refer to beer on occasion, but this is not the core purpose of their Twitter presence. Indeed, while the network clusters themselves are clearly evident, most are also far from distinctly separated from one another, suggesting a considerable overlap in membership. Notably, a range of accounts representing key utilities and services organisations (from @abcnews through @Telstra to @Qantas) are found at the very centre of the graph, indicating that accounts from across the map are equally likely to connect with them.

In itself, the map provides an argument against suggestions that online communication as such and social media in particular must necessarily lead to a fragmentation of the public sphere into separate interest groups

which act as sealed echo chambers; the considerable interconnection between virtually all major clusters shows that there are substantial overlaps between almost all groups represented here, but also indicates that certain combinations are more likely than others: politics and business, for example, go together more readily than religion and lifestyle pursuits. Notably, only two geographically based clusters can be readily identified, pointing to an overall lack of regional divisions in the Australian Twittersphere; the two regional clusters (centred around Perth and Adelaide) appear at least in part to be related to strong uptake of Twitter by specific local business interests (public relations for Perth, tourism and wine for Adelaide), and may point to deliberate strategies in these industries to use Twitter to strengthen informal local networks. For the vast majority of users which appear on the map, and for the clusters they form around shared interests, the interlinkages across communities are strong and widespread.

Several notable minor clusters can be identified at a distance from the main map, including various flavours of Christianity, teen culture, and education. The fact that these clusters are separate from the mainstream Australian Twitter network is likely due to their stronger orientation towards international interests which are not represented on this Australian map: Australian fans of teen stars Justin Bieber or the Jonas Brothers are more likely to use Twitter to connect with their peers elsewhere in the world than with other Australian users; Australian evangelical Christians will network with fellow believers in the United States and Europe. Although they may appear isolated on the map in Figure 1, then, they may nonetheless be part of rich Twitter networks which exist beyond this specific geographical space.

Networks in Action

These new data on the shape and extent of the overall Australian Twitter userbase, which Figure 1 represents, may now be used to shed new light on the Twitter user activities which are described by hashtag datasets. Where previously, it was possible to determine the total volume of activity within a hashtag, or to count the total number of unique users participating in the dataset, we are now able to develop a considerably more detailed picture of the spread of a hashtag, both in total and over time. Figure 3 illustrates these opportunities by comparing participation patterns in the #auspol and #ausvotes hashtags (in darker colours), against the backdrop of the overall map (in light grey): participants in #auspol are recruited largely from that part of the overall network which we have previously identified as being centred around politics, news, and journalism, and the level of participation in the hashtag is strongest in a small cluster to the left of the network which closer qualitative analysis of key accounts and their tweets reveals to be dominated by particularly hardline conservative views (one such user threatens in his profile description that he will block any left-winger who tries to follow his tweets, for example). Participation in #ausvotes during the 2010 election campaign, on the other hand, while still especially strong in the news and politics area of the overall map, is distributed considerably more evenly across all areas of the network; this is due perhaps to the fact that compulsory voting requires all eligible Australians to participate in the election at least to a minimal extent, as well as to the fact that Twitter was also used as an important backchannel to election night television coverage (Burgess & Bruns 2011) – both these factors appear likely to engender a more widespread use of the #ausvotes hashtag beyond established “political junkies” (Coleman 2006). Notably, activity by users in the hardline conservative cluster is substantially quieter in #ausvotes than it is in #auspol.

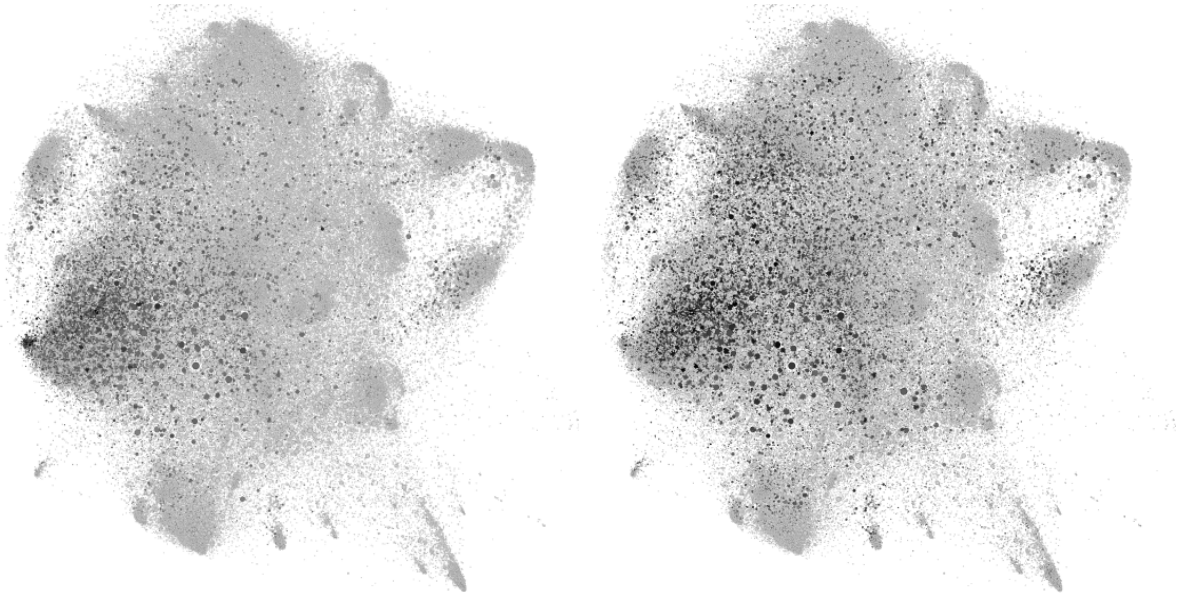


Figure. 3: Participation by Australian users in the hashtags #auspol (left) and #ausvotes (2010; right)

Further analysis, which is beyond the scope of this article, can further break down these overall patterns, for example by distinguishing the various phases of the election. It is able to document the differences in widespread popular engagement in the #ausvotes hashtag, between the early days of the election campaign and the final event of election night itself, or to show how actively different participants in #auspol react to the events of the day. Similarly, by using hashtag data to show which users tweet it at what point, it is possible to trace how specific information (for example, links to new articles about political matters, retweets, or other viral memes) is disseminated across the network over time. A useful metaphor to explore the potential of such approaches is the brain scan: it can be used to show how specific electrical impulses (tweets) traverse the synaptic structures of the brain (the Australian Twitter network) in response to various external stimuli.

While hashtags such as #auspol and #ausvotes are necessarily centred mainly around already identified thematic clusters in the overall network, other hashtags behave considerably differently; this is true for example for major televised events which draw participation from a wide range of the Twitter audience (and increasingly incorporate such backchannel activities deliberately; cf. Harrington *et al*., 2012). Figure 4 shows the user activity patterns in the hashtags for popular Australian cooking show *Masterchef* (#masterchef), and for the 2011 wedding between Prince William and Catherine Middleton (#royalwedding). These examples indicate substantial levels of activity across the overall map, pointing to the fact that these events and their coverage reach mass audiences rather than specific interest groups.

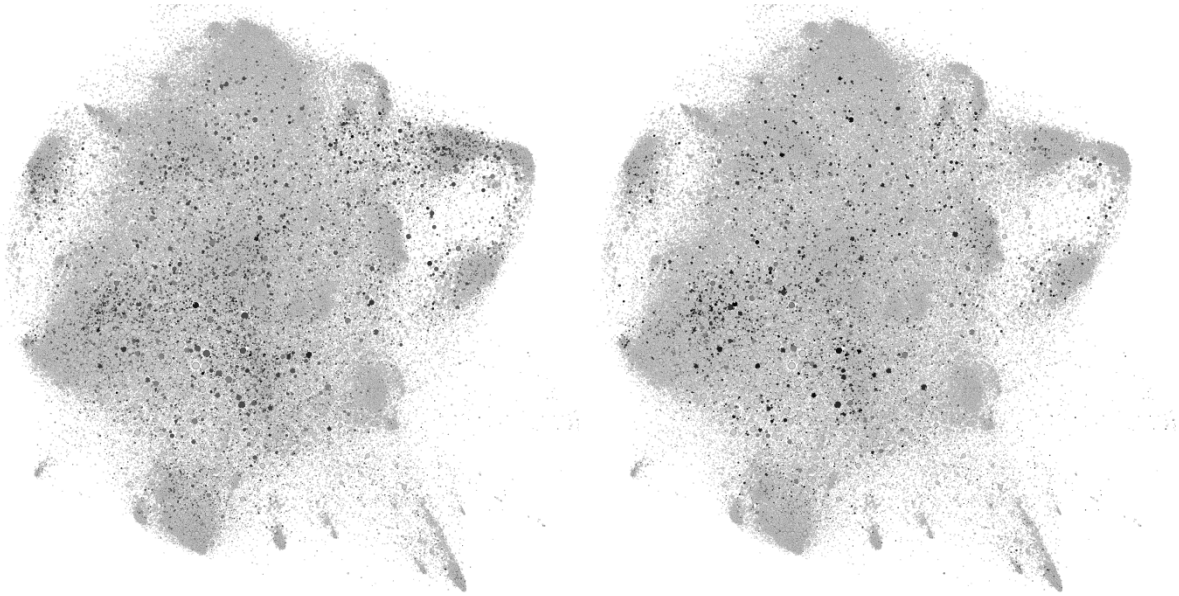


Figure 4: Participation by Australian users in the hashtags #masterchef (left) and #royalwedding (right)

This investigation of both the depth (volume) and breadth (network spread) of Twitter participation, then, also has immediate practical value for media researchers and media organisations; it provides a useful indication of the specific target audiences for diverse media events, and of the broadcasts' ability to engage with and enlist such audiences in active participation. Such research extends well beyond the hashtag datasets which we have explored so far: instead, it also becomes possible to examine the spread of key terms, of references to specific organisations or individuals, and of links to particular Websites (at domain or URL level). This style of analysis also has substantial applications in marketing and brand communication, therefore.

To illustrate this potential, Figure 5 compares the sharing of links to the Websites of the Australian Broadcasting Corporation (abc.net.au; *ABC News* sections only) and middle-of-the-road news site *news.com.au*, including in tweets where such URLs were shortened using *t.co*, *bit.ly*, or other URL shorteners, during February 2013. This shows the relative distribution of audiences for both sites across the overall network and its interest-based clusters: as one of the leading providers of news on state and federal politics in Australia, *ABC News* has a significant footprint especially amongst the politics-related clusters in the network; *news.com.au*, by contrast, has a less central role in such discussions and instead provides a greater range of general, entertainment, and sports news, resulting in a more scattered presence throughout the map. This indicates the relative strengths and weaknesses of both news organisations across the fields they cover, and can be used to inform the further strategic positioning of both sites.

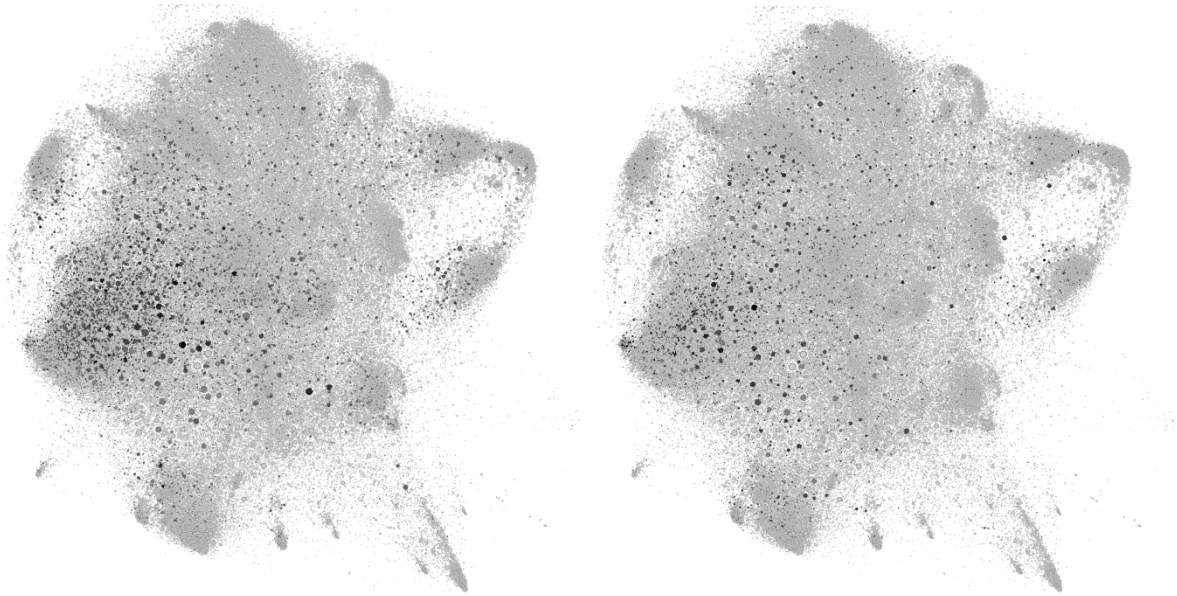


Figure 5: Sharing of links to *ABC News* and *news.com.au* by Australian users

Available space in this chapter does not permit us to outline a range of other possibilities for combining existing datasets with the network of follower/followee connections in the Australian Twittersphere – but at the same time, network maps such as this are rarely an end in itself, and rather constitute a means to an end. In the present case, for example, the visualisation of Twitter activity across the network map serves as a device to generate further research questions and challenges: once the different clusters have been identified, for example, it becomes possible to compare and contrast their specific compositions (how inward-looking or externally connected are they; how central are their best-connected nodes?), or to explore differences in the specific Twitter activities of the members of different clusters (do different clusters tweet more or less actively than others; do they prefer different news sources?).

Beyond Hashtags, Towards 'Big Data'

Beyond the map itself, the identification of a large part of the total Australian Twitter population also enables the development of entirely new models of gathering Twitter data, in addition to topic-based, "hashtag" studies. At least in theory, the database of Australian users which we have gathered would also enable us to track the public communication activities of all (or of a selected subset) of these users, to develop a more heterogeneous understanding of the day-to-day activities, interests, and contributions of Australian Twitter users. This substantially more comprehensive dataset would not need to rely on the topic selection process required to collect hashtagged tweets, and in particular it might cast important new light on the thematic and social dynamics of everyday, 'ordinary' tweeting outside of the limelight of the kinds of public events (elections, uprisings, Olympic games) framed as significant by the existing news values of the traditional media.

Such even more data-intensive, and arguably also even more data-*driven* research (which gathers Twitter data at large scale and on an ongoing basis in order to identify and explore activity patterns, ahead of formulating specific hypotheses to be addressed in the research) is possible in principle, but severely hampered in practice by the increasingly restrictive nature of API-based access to Twitter data (see Puschmann & Burgess 2013). While the tracking and capture of hashtagged tweets remains comparatively trivial and unencumbered, Twitter requires researchers wishing to track more than 5,000 active users at any one time to purchase data access from its commercial reseller Gnip, at volume-dependent costs which are likely to remain unaffordable for publicly-funded projects or individual researchers. Truly 'big data' research into the use of Twitter (and similar social media platforms) for the most part remains the realm of commercial market research, therefore,

limiting the availability of reliable and verifiable information on the diverse roles of such social media platforms in contemporary society (cf. Burgess & Bruns 2012b). Considering the growing role of social media in public, private, and "privately public" (Papacharissi 2010; also cf. Schmidt 2011) communication, this lack of researcher access to such important communicative resources is deeply problematic.

However, even if and when such access is available, there remain considerable other challenges in the further development of the digital humanities. First, many of the tools and technologies required to work with such 'big data' resources remain in early stages of their development, and there is a tendency for the simultaneous development of similar tools in separate research facilities which remain unaware of one another's efforts. Our own research project has chosen to make the research tools we have developed publicly available under Creative Commons licences wherever possible, therefore, and to similarly document our methodologies in significant detail through the project's Website at http://mappingonlinepublics.net/. Further, such methodological innovation takes place under precarious conditions, as continuing changes to the affordances and functionality of the Twitter API document: changes to API request and data formats, and to the rules and conditions of access, may undermine once promising research approaches at very short notice, and without an opportunity for appeal or renegotiation.

Finally, both the development and the use of such research methodologies and tools also requires a combination of disciplinary expertise which is not readily found in most humanities researchers, nor developed in mainstream research training. Working with 'big data' necessarily requires a certain degree of mathematical and statistical knowledge – not commonly a strength of media, cultural, and communication studies graduates; social media APIs and the tools to use them also necessitate considerable technical expertise. There is a considerable risk that mathematical and technical tasks are treated as 'black box' activities which may be delegated to research assistants and developers but are never questioned or problematised; this limits the range of interpretive activities which are possible, as well as hindering further methodological and conceptual innovation.

Finally, we are acutely aware of the limitations to our own ability to present a full discussion of our approaches to visualising the Twitter network maps we have presented above, and their implications, not for the straightforward representation of actually-existing networked publics, but in fact for the production of new 'calculated publics' (Gillespie, 2013, n.p.). In addition to the choices we have made about the size and composition of the original datasets (such as focusing on the most connected Australian accounts), each visualisation involves a substantial number of choices about visualisation algorithms and their specific settings. Each of these choices would have produced different versions of the 'calculated public' we are calling 'the Australian Twittersphere', but a detailed critique of these choices would not have been possible in the space of this chapter. Indeed, we have gone dangerously close to treating the data visualisation process as a 'black box', where network data are fed in at one end, and a network graph is spat out at the other. To highlight these limitations will at least serve as a reminder of the performativity of digital methods and an acknowledgement that there is far more work to do as we further develop the intersection of social media and digital humanities via data-driven research methods and practices—through reflexive critique, ongoing research training and interdisciplinary engagement we will need to continue working at cracking open and reassembling the black boxes that are the very stuff of the 'computational turn' (Berry 2012).

References

Berry, David. “The Computational Turn: Thinking about the Digital Humanities.” *Culture Machine* 12 (2011).

Berry, David, ed. *Understanding Digital Humanities*. London: Palgrave Macmillan, 2012.

boyd, danah, Scott Golder, and Gilad Lotan. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter." HICSS-43, Kauai, Hawai’i, 2010. http://www.danah.org/papers/TweetTweetRetweet.pdf.

boyd, danah. “Social network sites as networked publics: Affordances, dynamics, and implications.” In *Networked Self: Identity, Community, and Culture on Social Network Sites*, edited by Zizi Papacharissi, 39-58. New York: Routledge, 2010.

Bruns, Axel, and Jean Burgess. "#ausvotes: How Twitter Covered the 2010 Australian Federal Election." *Communication, Politics & Culture* 44, no. 2 (2011): 37-56.

Bruns, Axel, and Hallvard Moe. "Structural Layers of Communication on Twitter." In *Twitter and Society*, edited by Katrin Weller *et al*. New York: Peter Lang, forthcoming 2013.

Bruns, Axel, and Stefan Stieglitz. “Quantitative Approaches to Comparing Communication Patterns on Twitter.” *Journal of Technology in Human Services* 30, no. 3-4 (July 2012): 160–185. doi:10.1080/15228835.2012.744249. http://www.tandfonline.com/doi/abs/10.1080/15228835.2012.744249.

Bruns, Axel, and Stefan Stieglitz. “Towards More Systematic Twitter Analysis: Metrics for Tweeting Activities.” *International Journal of Social Research Methodology*, 22 Jan. 2013. DOI: 10.1080/13645579.2012.756095.

Bruns, Axel, Jean Burgess, Kate Crawford, and Frances Shaw. *#qldfloods and @QPSMedia: Crisis communication on Twitter in the 2011 South East Queensland floods*. Brisbane: ARC Centre of Excellence for Creative Industries and Innovation, 2012. http://cci.edu.au/floodsreport.pdf.

Bull, Tim. “How Many Australian Twitter Users Are There, and Where Are They From?” *Tribalytic Blog*, 2010. http://blog.tribalytic.com/how-many-australian-twitter- users-are-there-and-where-are-they-from/.

Burgess, Jean, and Axel Bruns. “(Not) The Twitter Election: The Dynamics of the #ausvotes Conversation in Relation to the Australian Media Ecology.” *Journalism Practice* 6, no. 3 (2012a): 384–402. doi:10.1080/17512786.2012.663610.

Burgess, Jean, and Axel Bruns. “Twitter Archives and the Challenges of ‘Big Social Data’ for Media and Communication Research.” *M/C Journal* 15, no. 5 (2012b). http://journal.media-culture.org.au/index.php/mcjournal/article/view/561.

Burns, Alex, and Ben Eltham. “Twitter Free Iran: An Evalution of Twitter’s Role in Public Diplomacy and Information Operations in Iran's 2009 Election Crisis.” *Record of the Communications Policy & Research Forum 2009* (2009): 298–310.

Christensen, Christian. "Twitter Revolutions? Addressing Social Media and Dissent." *Communication Review* 14, no. 3 (2011): 155-157.

Coleman, Stephen. “How the Other Half Votes: Big Brother Viewers and the 2005 General Election.” *International Journal of Cultural Studies* 9, no. 4 (2006): 457–479. doi:10.1177/1367877906069895.

Deller, Ruth. "Twittering On: Audience Research and Participation Using Twitter." *Participations* 8, no. 1 (2011). http://www.participations.org/Volume%208/Issue%201/deller.htm.

Dröge, Evelyn, Parinaz Maghferat, Cornelius Puschmann, Julia Verbina, and Katrin Weller. "Konferenz-Tweets: Ein Ansatz zur Analyse der Twitter-Kommunikation bei wissenschaftlichen Konferenzen." *Proceedings of the 12th International Symposium for Information Science*. Verlag Werner Hülsbusch, 2011. http://ynada.com/pubs/isi2010.pdf.

Gephi.org. “Gephi, an Open Source Graph Visualization and Manipulation Software.” 2012. http://gephi.org/.

Gillespie, Tarleton. “The Relevance of Algorithms.” In *Media Technologies*, edited by Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot. Cambridge, MA: MIT Press, 2013.

Harrington, Stephen, Tim Highfield, and Axel Bruns. "More than a Backchannel: Twitter and Television." In *Audience Interactivity and Participation*, edited by José Manuel Noguera. Brussels, Belgium: COST Action Transforming Audiences, Transforming Societies, 2012: 13-17. http://www.cost-transforming-audiences.eu/system/files/essays-and-interview-essays-18-06-12.pdf.

Highfield, Tim, Stephen Harrington, and Axel Bruns. “Twitter as a Technology for Audiencing and Fandom: The #Eurovision Phenomenon.” *Information, Communication & Society*, 3 Jan. 2013. DOI: 10.1080/1369118X.2012.756053.

Hartley, John. "From Cultural Studies to Cultural Science." *Cultural Science* 2, no. 1 (2009). http://www.cultural-science.org/journal/index.php/culturalscience/article/viewArticle/19/68.

Jacomy, Mathieu, Sebastien Heymann, Tommaso Venturini, and Mathieu Bastian. “ForceAtlas2, A Graph Layout Algorithm for Handy Network Visualization.” Discussion paper, 29 Aug. 2011. http://webatlas.fr/tempshare/ForceAtlas2_Paper.pdf

Larsson, Anders Olof, and Hallvard Moe. "Studying Political Microblogging: Twitter Users in the 2010 Swedish Election Campaign." *New Media & Society* 14, no. 5 (2011): 729-747.

Lotan, Gilad, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and danah boyd. "The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions." *International Journal of Communication* 5 (2011): 1375-1405.

Manovich, Lev. “Trending: The Promises and the Challenges of Big Social Data.” In *Debates in the Digital Humanities*, edited by Matthew K. Gold. Minneapolis: U of Minnesota P, 2012. 460-75.

Marwick, Alice E., and danah boyd. "I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. *New Media & Society* 13, no. 1 (2011): 114-133.

Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo. "Twitter under Crisis: Can We Trust What We RT?" 1st Workshop on Social Media Analytics (SOMA ’10), Washington, DC, 2010.

Palen, Leysia, Kate Starbird, Sarah Vieweg, and Anabda Hughes. "Twitter-Based Information Distribution during the 2009 Red River Valley Flood Threat." *Bulletin of the American Society for Information Science and Technology* 36, no. 5 (2010): 13-17.

Papacharissi, Zizi A. *A Private Sphere: Democracy in a Digital Age*. Cambridge: Polity, 2010.

———."Conclusion: A Networked Self." In *A Networked Self: Identity, Community and Culture on Social Network Sites*, edited by Zizi Papacharissi. New York: Routledge, 2011. 304-318.

Puschmann, Cornelius, and Jean Burgess. “The Politics of Twitter Data.” *HIIG Discussion Paper Series* 2013-01 (2013). http://dx.doi.org/10.2139/ssrn.2206225.

Schmidt, Jan-Hinrik. *Das neue Netz: Merkmale, Praktiken und Folgen des Web 2.0*. 2nd ed. Konstanz: UVK, 2011.

Stieglitz, Stefan, and Linh Dang-Xuan. "Social Media and Political Communication – A Social Media Analytics Framework." *Social Network Analysis and Mining*. Vienna: Springer, 2012. 1-15. DOI: 10.1007/s13278-012-0079-3.

Twitter. “Olympic (and Twitter) Records.” *Twitter Blog*, 12 August 2012a. http://blog.twitter.com/2012/08/olympic-and-twitter-records.html.

———. “GET Users/lookup.” *Twitter Developers*. 2012b. https://dev.twitter.com/docs/api/1.1/get/users/lookup.

Weller, Katrin, Evelyn Dröge, and Cornelius Puschmann. "Citation Analysis in Twitter: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences." #MSM2011 – 1st Workshop on Making Sense of Microposts, 2011. http://files.ynada.com/papers/msm2011.pdf.

Wilken, Rowan. "Twitter and Geographical Location." In *Twitter and Society*, edited by K. Weller *et al*. New York: Peter Lang, forthcoming 2013.