Axel Bruns[1], Stefan Stieglitz[2]

Twitter Data: What Do They Represent?

Abstract: Scholarly research into the uses of social media has become a major area of growth in recent years, as the adoption of social media for public communication itself has continued apace. While social media platforms provide ready avenues for data access through their Application Programming interfaces, it is increasingly important to think through exactly what these data represent, and what conclusions about the role of social media in society the research which is based on such data therefore enables. This article explores these issues especially for one of the currently leading social media platforms: *Twitter*.

ACM CCS: Information systems → World Wide Web → Web applications → Social networks

ACM CCS: Human-centred computing → Collaborative and social computing → Collaborative and social computing theory, concepts and paradigms → Social media

ACM CCS: Human-centred computing → Collaborative and social computing → Collaborative and social computing systems and tools → Social networking sites

Keywords: Twitter, Social Media, Internet Studies, Research Methods, Media Ecology

[1] **Corresponding Author: Axel Bruns**, ARC Centre of Excellence for Creative Industries and Innovation, Queensland University of Technology, Brisbane, Australia, eMail: a.bruns@qut.edu.au

[2] **Stefan Stieglitz**, Department of Information Systems, University of Münster, Germany, e-Mail: stefan.stieglitz@uni-muenster.de

1 Introduction: Social Media Data

Social media are now well-established as important platforms for everyday public communication, enabling a broader range of participants from ordinary citizens to cultural, economic, and political leaders to engage in public debate. One of the leading international social media platforms, *Twitter*, now boasts 271 million unique active users per month, from a total userbase of more than 750 million registered accounts.[1] Take-up of *Twitter* as a platform of communication remains unevenly distributed across and within societies, however: in countries such as the United States and Australia, for example, a substantial percentage of the population now has *Twitter* accounts, while take-up in Germany and Austria lags behind. Additionally, the demographics of the *Twitter* userbase in each country vary widely and reliable statistics are rarely available (e.g., following Globalwebindex, the penetration rate in the US is 20%, in Australia 12%, and in Germany 6%).[2]

Social media data – that is, data on the communicative activities of social media users, usually accessed through a range of Application Programming Interfaces (APIs) – have become increasingly important to researchers as sources of detailed, close to real-time information on the public response to current events, as part of an overall "computational turn" [1] towards 'big data' research in these fields. Although not without its pitfalls – see boyd & Crawford [3] for a useful discussion of the key challenges in working with 'big data' – the hope is that such approaches enable researchers to do more than merely study the Internet and its constituent platforms themselves: that they may allow us, instead, to study society with the Internet, as Rogers [26] has put it, by investigating how broader societal concerns are echoed in online communication.

In this context, then, and given the uneven distribution of participation in specific social media platforms as well as the comparative novelty and variety of the APIs which provide access to data on how participants are utilising these platforms [23, 24], it becomes crucially important to reflect on what these data represent, and to what extent they may therefore be relied upon as mirrors of society itself. This must address two key aspects: first, whether and how the communicative data available for any one social media platform represent the full breadth of communicative activities taking place on the platform itself (that is, how well the platform is able to represent itself), and second, to what extent these platform-specific data represent overall public debate (that is, how well the platform is able to represent society). While these questions can and must be asked of each online and social media platform that is of interest to researchers, in this article we focus specifically on *Twitter* as one of the most prominent platforms of the current generation of social media.

2 Do *Twitter* Data Represent *Twitter*?

The first question seems easy to answer: of course the data on users' communicative activities which are available through the APIs represent at least some of what happens on *Twitter* – however, the restrictions imposed by Twitter, Inc. on the uses of its public APIs combine with the operational limitations of most research projects to promote certain research approaches over others, almost independently of a specific method's fitness for the intended purpose.

Twitter, Inc. has gradually implemented a two-class data access regime in which a limited volume of data remains available through the standard, freely available APIs, while more comprehensive, higher-volume access is accessible only from commercial data resellers such as *Gnip* (recently bought by Twitter, Inc.) and *DataSift*, placing such data out of reach of much scholarly research due to the costs involved. The limited availability of standardised tools for gathering and analysing *Twitter* data has also meant that scholarly *Twitter* analysis has focussed on a handful of aspects of *Twitter* activity, which current research methods and tools can capture and investigate comparatively easily, while others have been largely ignored. The majority of such research has been forced to focus on the low-hanging fruits, and has struggled with the implementation of more complex and sophisticated research agendas [32].

For example, much early *Twitter* research has investigated patterns of communication within specific hashtags, for obvious practical reasons. Hashtag research provides an opportunity to examine how communities gather and interact around shared topics from politics [4, 20, 21, 38] and brands [19] through crises [10, 25, 33, 36, 37] to entertainment [13, 16, 17], and comparisons of activity patterns across hashtags are also possible, though still scarce [7, 8, 30].

But hashtag-centric studies must necessarily struggle to fully represent public communication on *Twitter*, even around the themes and topics the hashtag itself refers to. First, data access to the free *Twitter* API is throttled to a maximum of one per cent of the total current *Twitter* volume (if the current global activity on *Twitter* is 200,000 tweets per minute, the API will deliver a maximum of 2,000 tweets per minute for the

[1] 271m: https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=862505 (2014-07-30) / 750m: data gathered by authors

[2] http://blog.globalwebindex.net/twitter-the-fastest-growing-social-platform-infographic/ (2014-04-20)

search terms currently being tracked by an API user). Thus, unless researchers can afford to pay commercial rates, hashtag-based research into ‘large’ events will be severely limited in its accuracy.

Second, hashtag research depends crucially on the existence of a widely adopted hashtag, and on its (early) detection and tracking by researchers: where multiple alternative hashtags are being used, where researchers tracked a less widely used hashtag but missed out on a more popular alternative, or where hashtag use for public discussion of specific issues is not widespread, the data which can be gathered by tracking selected hashtags or even keywords will not represent the full breadth of relevant discussion.

Finally, and most crucially, hashtags represent only a self-selecting fraction of tweets (and users), missing out on an unknown volume of content which may relate to the same issues, but did not contain any relevant textual markers. During the 2011 Japanese tsunami, we identified four times as many tweets containing ‘tsunami’ than ‘#tsunami’; many more relevant tweets would not have used either term. Even many of the tweets responding to hashtagged messages do not themselves contain hashtags; hashtag-only datasets miss out on such follow-on communication. Further, those accounts that use hashtags regularly may be “*Twitter* experts”, and different from other users in their behaviour and activity. We found the #auspol hashtag (for Australian politics) to be dominated by a very small number of extremely active users, for example [8]. In this sense, hashtag-based datasets might not be representative for overall *Twitter* communication.

If hashtag datasets represent only the (self-selecting) tip of the iceberg of discussion on *Twitter*, they are unable to represent anything but a small part of *Twitter* activity. This small part is often valuable and interesting in its own right, precisely because of its self-selecting nature; however, the analysis of these data must be complemented by other approaches which are able to shed light on different aspects of the uses of Twitter for public communication.

As Bruns & Moe [6] show, hashtagged interaction forms one of three key layers of communication on *Twitter*. In addition to this macro-layer, which enables the rapid formation of ad hoc publics [5], there is also a meso-layer of everyday communication across the follower networks of individual *Twitter* accounts [31]. The bulk of *Twitter* activity takes place here, as users post tweets which are visible to their “personal publics” [27] of followers and are passed on by these followers through retweets. Further, beyond this stochastic distribution of content, which depends on followers happening to check their *Twitter* feeds at the time that new tweets are posted in their networks, a third, micro-layer of more direct but still public interaction is constituted through the exchange of @replies.

These other layers of interaction require different approaches to data gathering and analysis, which to date are less developed than hashtag analytics. To trace the distribution of a user’s tweets across the network, it would be necessary to map the network (to gather information on the user’s followers and followees, and for each of these connections, iterating the process over several steps). API restrictions imposed by Twitter, Inc. make this a very drawn-out process, which has been attempted only rarely by researchers (but see Bruns et al. [9] for one such initiative).

To explore the everyday communicative activities of selected users, outside of specific hashtags, it would be necessary to track all of the public tweets they send and receive, which is possible in a relatively straightforward manner through the free APIs – but to do so in a way that enables researchers to detect general patterns of activity beyond individual idiosyncrasies would require the tracking of large numbers of subjects, which again is likely to trigger API limits. An alternative approach, again likely to require funding to purchase access, is to connect to the sample streams of *Twitter* activity that Twitter, Inc. makes available through its APIs, from the ‘Spritzer’ (a random selection of one per cent of the total volume of current tweets) to the ‘Firehose’ (a comprehensive feed of all incoming tweets at any one moment); Gerlitz & Rieder [15] discuss whether sample feeds such as the ‘Spritzer’ can be considered to be representative of global *Twitter* activity.

The general lack of more comprehensive work that focusses especially on the meso- and micro-layers of *Twitter* communication also means that much of the research into hashtags has remained comparatively isolated. It is impossible to fully evaluate the relevance and impact of specific hashtagged discussions without being able to locate them in a wider communicative context. Did the participants in a prominent hashtag all belong to established networks of mutual follower relationships (were they the ‘usual suspects’ who always discuss this topic), or did the hashtag draw on participants from further afield (did it break out beyond an established niche group of interested users)? Does the volume of activity in a given hashtag constitute a large or small percentage of the total *Twitter* activity within a given population of users, however defined (was it one of many concurrent topics, or did it dominate discussion at the time)? It is only if such questions can be answered that we are able to fully assess what our *Twitter* data actually represent, both in terms of public discussion on *Twitter* and in terms of public debate in society as such.

3 Does *Twitter* Represent Society?

Given this lack of a comprehensive perspective on what forms of public communication take place on *Twitter* itself, it may seem premature to ask how well *Twitter* activities reflect societal concerns. But given that researchers as well as popular media are already positioning *Twitter* and other social media platforms as a window on society itself [2, 31], it is important to begin to address such questions: even where no comprehensive perspective of *Twitter* activities exists, more specific *Twitter* phenomena are now being used to trace wider societal patterns. Bruns *et al*. [11], for example, outline the Australian *Twitter* News Index (ATNIX), which traces the sharing of links to a selection of key Australian online news sources to provide an indication of which news stories are currently driving public debate. The US Geological Survey has begun to use *Twitter* as a human sensor network to complement its seismic sensors for the detection of earthquakes, apparently with a high degree of accuracy and at speeds which are limited by the distribution of tweets through the network rather than by the transmission of soundwaves through rock [14].

Such initiatives generate valuable information, but must be assessed against what is known about *Twitter*'s societal and geographic spread. ATNIX shows, in the first place, what news articles Australian *Twitter* users have seen fit to share with their networks (not always what they have read or what they agree with), and is influenced by the demographics of *Twitter* in Australia and by the demographics of the subsets of the Australian *Twitter* userbase that see sharing news links as part of their personas on *Twitter*. It cannot be understood as an uncomplicated reflection of the news interests of everyday Australians, or even of those Australians who have *Twitter* accounts. Similarly, USGS *Twitter* earthquake observations do not map easily onto seismic observations: they are subject to geographic variations in population density as well as *Twitter* take-up across the US (with the coastal population centres likely to be overrepresented in both), and to differences in earthquake sensitivity in these populations (earthquake-hardened Californians may react less vocally to an event of the same magnitude than residents of less tremor-prone regions).

Patterns of societal activity observed through the lens of *Twitter* research are therefore dependent on a range of additional variables which must be examined and understood afresh for each case. There is considerable risk that observations for one region or country may be translated inappropriately to different research environments. In 2011, the Pew Project reported that *Twitter* take-up in the US was especially strong amongst adolescent and African-American users [29], but we cannot conclude from this that *Twitter* is popular amongst similar groups in other countries as well – in Australia, for example, adoption appears to be greatest within a 25-55-year-old demographic that shares few traits with its US counterparts [28]. Any direct translation of analytical frameworks (or commercial strategies) from one context to the other will miss its mark, therefore.

At the same time, limited take-up of *Twitter* as a communications tool in any one country does not necessarily translate into limited relevance in public debate; Australia shows adoption particularly by an especially influential societal group, and research in other countries appears to indicate similar patterns (see e.g. Maireder & Ausserhofer's study of *Twitter* in Austrian politics [22]). If *Twitter* represents especially the activities of societal opinion leaders such as journalists or politicians [18], then it may well wield influence beyond its observable market share.

This, then, requires us first to understand *Twitter*'s role within the overall media ecology, before assessing what aspects of public debate it can represent, and how well it is able to do so. Fully realized, this research agenda is likely to extend well beyond what can be examined through an analysis of 'big data' drawn from *Twitter* itself, expanding instead into studies which investigate the flows of information in society across multiple online and offline channels and platforms, and explore the individual media repertoires of different citizens. Against this background, it may then be possible to determine what part of the overall public debate is represented by activities on *Twitter* itself.

4 Conclusion

Even without such more comprehensive frameworks for the study of public debate, *Twitter* research can and does make important contributions to our understanding of public communication across society. The critical questions we have asked of hashtag-centric studies do not undermine the utility of such research initiatives, but do define some limits to the applicability of their findings; most of all, they encourage the further development of complementary research agendas which seek to capture some of the higher-hanging fruit in *Twitter* research [32], even in spite of the increasing number of obstacles placed in the path of such research activities by Twitter, Inc. itself (cf. [12]).

Although it is difficult, ground-breaking research which sheds a different light onto *Twitter* communication activities is underway in a number of research centres (including, for example, the Digital Media Initiative, University of Amsterdam [15, 26]; Microsoft Research New England, Boston [3, 21]; the

Social Media Lab, University of Washington [34, 35]; and the Social Media Research Group, Queensland University of Technology [7, 9]), often by developing longer-term strategic research agendas rather than engaging in more short-term, *ad hoc* research into specific momentary phenomena. Such strategic research should also increasingly seek to connect with contingent research activities in related fields outside of *Twitter* and social media studies themselves, to develop a broader, cross-platform perspective which is able to offer new perspectives on the contemporary media ecology itself. Finally, of course, usage practices as they relate to *Twitter,* social media, and the Internet as such also remain in constant flux, and any research in this field must remain agile enough to adjust to such changes. A lack of appropriate theories which advance beyond the mere explanation of data patterns might result in a superficial platform-centric review of data, without gaining higher-level knowledge. As a consequence, there is an urgent need to develop more sophisticated theories to interpret the data collected.

Literature

[1] D. Berry. The computational turn: thinking about the digital humanities. Culture Machine, 12, 2011.

[2] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. L. Ademic, R. Baeza-Yates, S. Counts (eds.), Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Palo Alto, CA: AAAI Press, 450-453, 2011.

[3] d. boyd and K. Crawford. Critical questions for big data. Information, Communication & Society, 15(5):662-679, 2012.

[4] A. Bruns and J. Burgess. #ausvotes: how Twitter covered the 2010 Australian federal election. Communication, Politics & Culture, 44(2):37-56, 2011.

[5] A. Bruns and J. Burgess. The use of Twitter hashtags in the formation of *ad hoc* publics. Proceedings of the European Consortium for Political Research conference, Reykjavík, 25-27 Aug., 2011.

[6] A. Bruns and H. Moe. Structural layers of communication on Twitter. K. Weller, A. Bruns, J. Burgess, M. Mahrt, C. Puschmann (eds.) Twitter and Society, New York: Peter Lang, USA, pp. 15-28, 2013.

[7] A. Bruns and S. Stieglitz. Quantitative approaches to comparing communication patterns on Twitter. Journal of Technology in Human Services, 30(3-4):160-185, 2012.

[8] A. Bruns and S. Stieglitz. Towards more systematic Twitter analysis: metrics for tweeting activities. International Journal of Social Research Methodology, 16(2):91-108, 2013.

[9] A. Bruns, J. Burgess, and T. Highfield. A ‘big data’ approach to mapping the Australian Twittersphere. K. Bode, P. Arthur (eds.) (Re)purposing the (Digital) Humanities: Research, Methods, Theories, Basingstoke: Palgrave Macmillan, forthcoming 2014.

[10] A. Bruns, J. Burgess, K. Crawford, and F. 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

[11] A. Bruns, T. Highfield, and S. Harrington. Sharing the news: dissemination of links to Australian news sites on Twitter. J. Gordon (ed.) Br(e)aking the News, New York: Peter Lang, 2013.

[12] J. Burgess and A. Bruns. Twitter archives and the challenges of “big social data” for media and communication research. M/C Journal, 15(5), 2012.

[13] R. Deller. Twittering on: audience research and participation using Twitter. Participations, 8(1), 2011.

[14] P. Earle, M. Guy, R. Buckmaster, C. Ostrum, S. Horvath, and A. Vaughan. OMG earthquake! Can twitter improve earthquake response? Seismological Research Letters, 8(12):246-251, 2010.

[15] C. Gerlitz and B. Rieder. Mining one percent of Twitter: collections, baselines, sampling. M/C Journal, 16(2), 2013.

[16] T. Highfield. Following the yellow jersey: tweeting the Tour de France. K. Weller, A. Bruns, J. Burgess, M. Mahrt, C. Puschmann (eds.) Twitter and Society, New York: Peter Lang, pp. 249-262, 2013.

[17] T. Highfield, S. Harrington, and A. Bruns. Twitter as a technology for audiencing and fandom: the #Eurovision phenomenon. Information, Communication & Society, 16(3):315–39, 2013.

[18] E. Katz and P.F. Lazarsfeld. Personal Influence: The Part Played by People in the Flow of Mass Communications. New York: Free Press, 1955.

[19] N. Krüger, S. Stieglitz, and T. Potthoff. Brand Communication in Twitter — a Case Study on Adidas. Proceedings of the Pacific Asia Conference on Information Systems (PACIS), Hochiminh City, Vietnam, Paper 161, 2012.

[20] A.O. Larsson and H. Moe. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media & Society, 14(5):729-747, 2011.

[21] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce and d. boyd. The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. International Journal of Communication 5:1375-1405, 2011.

[22] A. Maireder and J. Ausserhofer. Political discourses on Twitter: networking topics, objects, and people. K. Weller, A. Bruns, J. Burgess, M. Mahrt, C. Puschmann (eds.) Twitter and Society, New York: Peter Lang, pp. 305-318, 2013.

[23] F. Morstatter, J. Pfeffer, H. Liu, and K.M. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose, Proceedings of the International Conference on Weblogs and Social Media ICWSM, 2013.

[24] F. Morstatter, J. Pfeffer, and H. Liu. When is it Biased? Assessing the Representativeness of Twitter's Streaming API. Proceedings of the WWW Web Science - WWW’14 Companion, April 7–11, 2014, Seoul, Korea, 2014.

[25] L. Palen, K. Starbird, S. Vieweg, and A. Hughes. Twitter-based information distribution during the 2009 Red River Valley flood threat. Bulletin of the American Society for Information Science and Technology 36(5):13-17, 2010.

[26] R. Rogers. The end of the virtual. Inaugural address, University of Amsterdam. http://www.govcom.org/rogers_oratie.pdf

[27] J.-H. Schmidt. Twitter and the rise of personal publics. K. Weller, A. Bruns, J. Burgess, M. Mahrt, C. Puschmann (eds.) Twitter and Society, New York: Peter Lang, 2013, 3-14.

[28] Sensis. Yellow™ social media report: what Australian people and business are doing with social media. May 2013. p. 15. http://about.sensis.com.au/DownloadDocument.ashx?DocumentID=463

[29] A. Smith. Twitter Update 2011. Pew Internet and American Life Project, 1 June 2011. http://www.pewinternet.org/Reports/2011/Twitter-Update-2011/Main-Report.aspx (accessed 20th April 2014)

[30] M.A.Smith, L. Rainie, I. Himelboim, and B. Shneiderman. Mapping Twitter Topic Networks: From Polarized Crowds to Community Clusters. Pew Research Center, Feb. 20, 2014. http://www.pewinternet.org/files/2014/02/PIP_Mapping-Twitter-networks_022014.pdf (accessed 25 July 2014).

[31] S. Stieglitz and L. Dang-Xuan. Emotions and Information Diffusion in Social Media — Sentiment of Microblogs and Sharing Behavior. Journal of Management Information Systems, 29(4):217–248, 2013.

[32] S. Stieglitz, L. Dang-Xuan, A. Bruns, and C. Neuberger. Social Media Analytics: An Interdisciplinary Approach and Its Implications for Information Systems. Business and Information Systems Engineering 6(2):89–96, 2014.

[33] S. Stieglitz and N. Krüger. Analysis of sentiments in corporate Twitter communication – a case study on an issue of Toyota. Proceedings of the 22nd Australasian Conference on Information Systems, 2011.

[34] K. Driscoll and S. Walker. Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. International Journal of Communication, 8:1745–1764, 2014.

[35] K. Nahon, J. Hemsley, R. Mason, S. Walker, and J. Eckert. Information Flows in Events of Political Unrest, Proceedings of the iConference, Forth Worth, 2013. http://ekarine.org/wp-admin/pub/InformationFlowsInEvents.pdf (accessed 25 July 2014).

[36] O. Oh, M. Agrawal, and H. R. Rao. Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises. MIS Quarterly, 37(2):407-426, 2013.

[37] C. Ehnis and D. Bunker. The Impact of Disaster Typology on Social Media Use by Emergency Services Agencies: The Case of the Boston Marathon Bombing, Proceedings of the 24th Australasian Conference on Information Systems, Melbourne, Australia, 2013.

[38] R. Sandoval, R.T. Matus, and R.N. Rogel. Twitter in Mexican Politics: Messages to People or Candidates?, Proceedings of the Americas Conference on Information Systems, Paper 18, July 2012.

[39] Z. Hong, Z. Kem, M. Lee, and F. Feng. Enterprise Microblog as a New Marketing Strattegy for Companies: Enterprise Microblog Commitment and Brand Loyalty, Proceedings of the Pacific Asia Conference on Information Systems, Paper 74, 2013.

Dr Axel Bruns is an ARC Future Fellow and Professor in the Creative Industries Faculty at Queensland University of Technology, and a Chief Investigator in the ARC Centre of Excellence for Creative Industries and Innovation (http://cci.edu.au/). He is the author of *Blogs, Wikipedia, Second Life and Beyond* (2008) and *Gatewatching* (2005), and a co-editor of *Twitter and Society* (2013). For more details on his current social media research, see http://mappingonlinepublics.net/.

Address: ARC Centre of Excellence for Creative Industries and Innovation, Queensland University of Technology, Brisbane, Australia, eMail: a.bruns@qut.edu.au

Dr Stefan Stieglitz is an Assistant Professor at Department of Information Systems at the University of Münster in Germany. He is founder and academic director of the 'Competence Center Connected Organization'. His research focuses on economic, social, and technological aspects of collaboration software and has been published in reputable international journals such as Journal of Management Information Systems, Social Network Analysis and Mining, MIS Quarterly Executive, and International Journal of Social Research Methodology.

Address: Department of Information Systems, University of Münster, Germany, e-Mail: stefan.stieglitz@uni-muenster.de