Big Social Data Approaches in Internet Studies: The Case of Twitter

Prof. Axel Bruns
ARC Future Fellow
Digital Media Research Centre
Queensland University of Technology
Brisbane, Australia
a.bruns@qut.edu.au
@snurb_dot_info

Abstract

Well beyond Internet Studies itself, but arguably led by it to a considerable extent, there has been a turn towards computational methods in the study of social and communicative phenomena at large scale. This "computational turn" has commonly been described as a turn towards "big data" or, more specifically, towards "big social data", and it continues to drive the development of new research methodologies, approaches, and tools.

Internet Studies has been an advocate of "big data" approaches because the field connects several core disciplines that use "big data" methods – media, communication and cultural studies, the social sciences, and computer science. Equally, the major objects of research in Internet studies – including platforms, search engines, mobile apps and devices, and Internet technologies and networks themselves – are key sources of "big data" on user interests, attitudes, and activities. Proponents of such approaches suggest that it is becoming possible to "study society with the Internet", while others ask critical questions about which observations are privileged and which are discounted as the logic of "big data" influences research agendas.

The early development and application of "big social data" research methods in Internet Studies, as well as critical interrogations of such approaches, focused especially on research into *Twitter* as a global social media platform. This is largely due to *Twitter*'s (initially) highly accessible Application Programming Interface (API), which enabled the development of powerful research methods and the promise of large, sometimes real-time, datasets tracing patterns of user activity around specific themes and topics on the platform, as well as, by proxy, in wider society.

Twitter's tightening of API access serves as a reminder of the precarious nature of "big social data" research drawing on proprietary datasets, just as concerns about the use of social media data for the social profiling of individual users raise questions about research ethics and user privacy. The growing body of "big data" research drawing on *Twitter* as a data source has paradoxically also underlined the many limitations and blind spots of such approaches, as researchers drawing on publicly available API data struggle to place their findings in the context of a platform whose overall global shape is shrouded in considerably more mystery, due to Twitter, Inc.'s interest in keeping aspects of the platform and its user community commercial-in-confidence. The increased work in this field also highlights shortcomings in research training and publishing models, which need to be addressed to further develop "big social data" research.

This chapter outlines the current state of the art in computationally-driven Twitter research, using platform-specific research as a case study for the computational turn in Internet Studies. It will consider the opportunities and challenges inherent in this shift toward more data-driven research, and outline the key needs for the discipline which have emerged to date. Even as Twitter's own fortunes fluctuate, the experiences made in this branch of Internet Studies stand as a guide for broader developments in our field.

Introduction: 'Big Data' and the Computational Turn

Internet studies has always been a hybrid field which connects disciplines as diverse as cultural studies and computer science, and draws on methods ranging from ethnographic observation to social network analysis. This has been the cause of tensions and disconnects that have often also been highly productive, as they have required researchers from diverse disciplinary backgrounds to at least acquire a basic fluency in each other's languages in order to be intelligible to one another – but at times it has also furthered existing divisions between the disciplines that have kept them from developing a more collaborative approach to addressing common research questions and problems.

The importance of finding a common language between researchers and research groups who would variously describe themselves as more 'quantitative' or more 'qualitative', more 'computational' or more 'interpretive' in their underlying methodological orientation has grown further in recent years by the greater scientific and popular focus on 'big data', and such datasets' increasing relevance to the field of Internet studies. Digital, online processes are a particularly prominent generator of 'big data', as user activities in social media spaces, with mobile devices, or using the Internet in any other form (including the growing 'Internet of Things') are each leaving trails of data and metadata that are increasingly persistent. The existence of such highly detailed data trails, often at a resolution that enables the identification of individual users and devices, and the tracing of their activities on a second-by-second basis, has resulted in significant scholarly, commercial, and governmental interest in these 'big data' sources, variously aiming to develop better understandings of collective processes in society, more effective and personalised advertising and marketing mechanisms, or more comprehensive surveillance and intelligence systems. At the same time, substantial debate about the ethical and privacy implications of such 'big data'-enabled research, and about the general desirability of these developments, has also arisen (e.g. boyd & Crawford, 2012; Andrejevic, 2014), and we explore some of these questions in the discussion below.

The emergence of 'big data' introduces a number of key changes to the research process. First, the existence of more comprehensive datasets that appear to present whole-of-population patterns enables a move away from approaches that draw only on convenient or representative population samples constructed by the researcher; this, in turn, may also support a more finegrained analysis of minor patterns that may not have appeared clearly enough in such more limited samples. Second, any approach that seeks to work with entire, large datasets must necessarily confront new challenges in processing, analysing, and presenting the patterns observed in such datasets, drawing on advanced computational tools and techniques for data analytics and visualisation. Finally, this also complicates the presentation and critical evaluation of research outcomes in scholarly and other contexts, especially as far as peer review and research replicability are concerned, as academic peers rarely have access to the same datasets and analytical tools and may not yet have the data analysis skills required for following the discussion.

In a scholarly context, then, this emerging interest in doing research that incorporates the analysis of such very large and often dynamically growing real-time datasets has been described by Berry (2011) as a new "computational turn" especially in the humanities and social sciences, where it represents a marked departure from earlier approaches that had drawn only on considerably smaller datasets which did not require the use of computational methods for their analysis. To some extent this computational turn in what are now often being described as the "digital humanities" (cf. Arthur & Bode, 2014) can thus also be seen as an interdisciplinary turn, as it requires humanities researchers to connect with or at the very least learn from computer science in order to add further computational approaches to their conventional methodological toolkit; conversely, in doing so there may also be a corresponding "social turn" in computer science, as the analytical methods developed in that field are increasingly applied to real-world social science research problems. As a natural nexus between these disciplines, Internet studies is well positioned to facilitate and benefit from these convergent turns, as well as to critically examine the methodological and conceptual problems that may arise from them.

A particular driver of such attempts at disciplinary convergence is the area of research that draws centrally on what are sometimes described as 'big social data' (Manovich, 2012; Burgess & Bruns, 2012): the large-scale, real-time datasets on social interactions on the Internet – and particular in popular social media spaces such as *Facebook* and *Twitter*. Such datasets are able to trace in significant detail, and on an ongoing basis, the ways in which Internet users are engaging with and responding to the events taking place in the world that surrounds them, and proponents of such 'big social data'-driven research have therefore argued that their approaches may enable Internet studies to transition from "researching the Internet" to "studying culture and society *with the Internet*" (Rogers, 2009: 29) – that is, to use observations on patterns of online interaction as a lens through which to perceive society as such. Critiques of this view point out, however, that this lens is, at best, a flawed and distorted one, due to the various particularities and limitations of the datasets upon which it relies (cf. boyd & Crawford, 2012); it is therefore important at the very least not to lose sight of the specificities of the underlying datasets if any generalisation from the observable online to the more fundamental societal patterns is to be attempted.

This chapter, then, outlines the state of the art in contemporary *Twitter* research as a representative example of the broader computational turn and its implications for Internet studies. We explore especially the continuously evolving conditions for accessing and using 'big social data' from *Twitter*, and their implications for the conduct of rigorous and sustainable research into the uses of *Twitter* and their interrelationship with wider societal practices. We also examine the needs for digital methods training and methodological development that have emerged from the *Twitter* research experience over the past decade, and discuss the issues that arise from the precarious situation of working with datasets provided by a commercial entity whose politics and policies are shifting almost constantly. Independent of *Twitter*'s own further trajectory, the experiences made in this branch of Internet Studies stand as a guide for broader developments in our field.

Twitter as a Source of 'Big Social Data'

Twitter has become a particularly common example for the computational turn in Internet studies because of its traditionally relatively permissive approach to providing large datasets on user activities to scholarly and industry researchers, and because of the at least relatively limited ethical and privacy concerns in working with *Twitter* data. The overall network and communicative structure of *Twitter* is simple: users can choose between making their accounts globally public, which results in their profiles and posts being visible even to non-registered visitors to the *Twitter* Website, and setting their accounts to 'protected', which means that the full profile and its posts are visible only to the account's followers, and that these followers must be individually vetted and approved by the account holder before they gain access to the posts. As of September 2013, only some five per cent of all 843 million *Twitter* accounts were 'protected' in this way (Bruns et al. 2014a); the overwhelming majority of *Twitter* profiles and their posts are publicly available on the Web, and potentially indexed in various search engines. This is markedly different from the situation for other major social media platforms such as *Facebook*, then, where the vast majority of profiles and posts are visible only to approved 'friends' of a user, or where visibility can be adjusted on a post-by-post basis and may range from 'private' through 'friends only' and 'friends of friends' to 'public'.

Researchers may therefore generally infer that *Twitter* users understand that posts made by their non-'protected' *Twitter* accounts are publicly visible in the same way that material posted to other public Websites is publicly visible; this does not imply that such users are also necessarily aware of the potential for their posts to be included in data collections and analysed by scholarly or other researchers, however, and should therefore also not be seen as an implicit permission for researchers to publish research that contains detailed individual profiling of ordinary users' activities. However, as far as data collection itself is concerned, the very publicness of *Twitter* content is widely accepted by researchers, and by the ethics review boards which oversee their activities, to be sufficient support for the argument that gathering profile and post data from *Twitter* without the express personal consent of each user included in the collection is acceptable in terms of ethics and privacy.

Traditionally, such data collection from *Twitter* has been enabled by a comparatively open and powerful Application Programming Interface (API) that allowed large-scale collection of profile information and tweets.

The *Twitter* API is not primarily designed for research purposes, but supports a range of uses – most importantly, it enables the functionality of a range of third-party *Twitter* clients such as *Tweetdeck* (subsequently purchased by Twitter, Inc.) and *Hootsuite*, as well as of the various apps for smartphone and tablet devices which Twitter, Inc. itself provides. However, the API functions that enable such apps to search for specific keywords and hashtags, or to subscribe to a stream of updates from a given set of users or containing selected search terms are equally useful for researchers seeking to construct datasets that relate to specific events or user populations, and a range of *Twitter* data capture solutions for research and related purposes gradually emerged as *Twitter*'s growing popularity since its launch in March 2006 made it an increasingly interesting object of and environment for research.

While some of these solutions remained strictly in-house tools developed by specific research groups, others – such as the early standard *Twapperkeeper* – were set up as public services available to any researcher who registered on their Websites. Twitter, Inc. also explicitly supported some of these research initiatives by operating a 'whitelisting' scheme: while for the majority of API users, a range of access limits applied (restricting the maximum number of accounts or keywords that any one API client could follow, or the maximum throughput of content returned by the API), the standard limits could be eased for those developers and researchers who made an informal, well-justified request to the *Twitter* developer support team. This supportive, unbureaucratic access regime resulted in the emergence of a growing developer and researcher ecosystem around the fledgling social network as it evolved.

As a result, the field of *Twitter* research also flourished, covering an increasing breadth of topics that ranged from crisis communication (Hughes & Palen, 2009; Mendoza et al., 2009; Palen et al., 2010) through political discussion (Golbeck et al., 2010; Larsson & Moe, 2011; Vergeer & Hermans, 2013) to collective audiencing (Harrington et al., 2013; Highfield et al., 2013) and building on a variety of more and more sophisticated methods and tools that especially sought to establish more powerful frameworks for dealing computationally with near-live and increasingly large datasets. The widespread use of a handful of publicly available and/or open-source research tools also contributed to a gradual standardisation of basic analytics methods, enhancing the compatibility and comparability of methods and studies (Bruns & Stieglitz, 2013; 2012). Studies of the roles played by *Twitter* in the context of various major events – such as the 2009 Iranian presidential election (Gaffney, 2010; Rogers et al., 2009), 2010 Haiti earthquake (Bruno, 2011; Sarcevic et al., 2012), or 2011/12 Arab Spring uprisings (Lotan et al., 2011; Mourtada & Salem, 2011; Meraz & Papacharissi, 2013) also generated significant public attention, and provided Twitter, Inc. with valuable independent evidence for its growing role in public communication.

Not all such studies should necessarily be understood as drawing on 'big social data' in any meaningful definition of the term. Some built, for example, on datasets of hashtagged *Twitter* conversations containing no more than some thousands or tens of thousands of tweets – sizeable collections, certainly, but hardly of a magnitude that requires a radical shift towards entirely new computational methods designed specifically for very large datasets. (It should be noted in this context that the boundaries of what is 'big' in 'big data' have remained notoriously undefined – one not entirely tongue-in-cheek definition, however, is of 'big data' as anything that exceeds the maximum of 1,048,576 data rows that are allowable in current versions of Microsoft Excel.) However, these smaller-scale studies often proved to be the stepping stones towards much larger-scale research projects which rapidly exceeded the processing power of conventional research tools: from examining the role of *Twitter* in comparatively minor and localised natural disasters to tracking the millions of tweets responding to the March 2011 earthquake, tsunami, and still unresolved nuclear disaster on the Japanese east coast (Acar & Muraki, 2011; Doan et al., 2012); from exploring the 2009 presidential election in Iran to investigating its 2012 counterpart in the United States (Conway et al., 2013; Bruns & Highfield, forthcoming); from tracking audience engagement with a one-off media event to developing comprehensive longitudinal metrics for the use of *Twitter* as a backchannel for sharing and commenting on mainstream media content in an entire national mediaspace (Bruns et al., 2013; Bruns et al., 2014b). Here, new computational methods for data access, processing, storage and analytics quickly emerged as necessary requirements.

This growth in and broadening of research approaches in recent years reflects not only the increasing importance of *Twitter* in global public communication, but also the gradual realisation that earlier, more limited research approaches continued to have significant blind spots. This realisation is reflected especially in a gradual move beyond the initial overreliance on what may be described as “hashtag studies” (Burgess & Bruns, 2015). Early *Twitter* research, drawing on the most obvious affordances of the *Twitter* API and the available data capture tools, proceeded especially by capturing and analysing datasets that were defined by the presence of a specific thematic hashtag in the data – from terms such as #ausvotes for Australian federal elections (Bruns & Burgess, 2011) to #sandy for the 2012 hurricane on the U.S. east coast (Hughes et al., 2014). But such datasets necessarily cover only a self-selecting subset of all *Twitter* uses relating to the same events: they include only those tweets which their authors chose to mark as relevant to the topic, by including one particular hashtag. Exchanges using competing or complementary hashtags, or topically relevant but non-hashtagged conversations, remain excluded from these datasets unless researchers went to the trouble of tracking a variety of alternative hashtags and keywords, perhaps even updating that list of terms to be included in real time as the event unfolded and user practices evolved.

Such “hashtag studies”, even where they are expanded to cover multiple parallel hashtags and keywords, constitute only one approach to investigating the uses of *Twitter* in relation to specific themes, topics, and events, therefore. An alternative approach is to examine the activities by and around selected populations of *Twitter* accounts: for example by capturing all of the public tweets that originate from, or reference in the form of @mentions and retweets, a set of public figures, political actors, or random selection of ordinary users. Undertaken for example for all of the *Twitter* accounts operated by current Members of Parliament and parliamentary candidates in a national election (cf. Bruns, 2014), this is likely to shed a very different light on the visibility of and response to these various politicians, compared to an analysis of the self-selecting stream of hashtagged commentary on the election – and in comparison, the account-centric and hashtag-centric analyses may evaluate the extent to which the hashtag conversation is at all representative of the wider *Twitter* debate relating to the election. Similarly, through the *Twitter* search API it is also possible to capture the tweets which contain links to specific domains (even if those links have been shortened using *Twitter*’s mandatory URL shortener *t.co* and/or other shortening services); independent of selected hashtags and keywords, and of specific user populations, this can generate a comprehensive perspective on the dissemination of particular information across the Twittersphere (cf. Bruns et al., 2013; Bruns & Sauter, 2015).

Inter alia, it is thus possible to use the *Twitter* API to track and capture on a continuous basis the tweets that contain selected hashtags and keywords, are posted by or reference in their @mentions and retweets a selection of accounts, or link to particular content on the Web. In combination, this results in larger, hybrid, and more complex datasets that cover a broader range of user activities than the early “hashtag studies” alone had been able to do; this considerably expands the ability of *Twitter* research to explore how the platform is used. The datasets resulting from such approaches are likely to be substantially larger and more inclusive, forcing researchers to confront the computational and analytical challenges associated with ‘big social data’ at a scale that truly deserves this description. This places significant additional importance on the various software tools used for such work.

Doing ‘Big Social Data’ Research

Typical trends in data-driven *Twitter* research have been shaped in a number of ways by the research tools and platforms available to researchers. We have already noted *Twapperkeeper* as an early standard framework for gathering *Twitter* data in relation to hashtags, keywords, and other search terms – at first in its online incarnation as *Twapperkeeper.com*, and subsequently, after Twitter, Inc. forced the closure of this public service, as an open-source data gathering framework called *yourTwapperkeeper* which potential users had to install on their own servers. Notably, in both versions, *Twapperkeeper* only ever captured a subset of the full metadata delivered by the *Twitter* API alongside every tweet matching the tracking criteria set by the user; this selective retainment of the data also considerably influenced the direction of the research projects which utilised this tool, and influenced the development of the standard *Twitter* activity metrics which could be extracted from the data (Bruns & Stieglitz, 2013). Such divergences in the capabilities of data capture and management tools only emerge fully in comparisons between different research tools, however: at a time when (*your*)*Twapperkeeper* was by far the most prominent and most widely used tool for gathering data from *Twitter*, the vast majority of researchers treated it in essence as a neutral 'black box' through which the data stream passed on its way from the API to the researcher, but which needed little critical attention.

More recent, alternative frameworks for gathering data from the *Twitter* API, such as the *Twitter Capture and Analysis Toolset* (*TCAT*) developed by the Digital Media Initiative at the University of Amsterdam, have served to reveal in more detail some of the differing data gathering and processing approaches which are possible even while using the same *Twitter* APIs. *TCAT* pays considerably more attention to the limits of the *Twitter* API, for example: while using *Twapperkeeper* it is possible to add considerably more tracking terms than the free and open API is able to serve under current API access restrictions, resulting in the exclusion of an unknown volume of data from the datasets returned by the API, *TCAT* also captures the alerts which highlight that data are missing from the results returned by the API. This does not in itself enable the researcher to address these gaps in the data, but it at least creates an awareness that such gaps may exist, and enables the researcher to provide an estimate of how many tweets were missed. By contrast, similar gaps may be present in the datasets upon which much of the published research which used *Twapperkeeper* builds – but without the authors of such work being aware of, or able to quantify, the extent of these gaps.

Even as simple an example as this already highlights the need for Internet studies researchers to move beyond an understanding of their 'big data' research tools as unproblematic 'black boxes' which require no further discussion. In the first place, it is incumbent on any social media researcher to become intensely familiar with the functionality and limitations of the data access APIs they rely on, and of the data gathering tools they may use to connect to the API; further, especially in conducting 'big data'-driven research that relies inherently on computational methods for filtering, processing, analysing, and visualising the data it is also crucial for the researcher to examine very closely the operational assumptions embedded into the various softwares used at every step of the way. Unfortunately, there still persists in many published studies a tendency to skip a clear consideration of the processing choices made by researchers, especially prominently perhaps in discussing the data visualisation approaches employed.

As research moves further towards the use of truly 'big' datasets, which generate increasing data processing complexities and often require the development of custom-made analytics tools, such oversights become especially problematic, as they complicate the independent peer review and replicability of research results. But even before the research advances to a publishable state, there is a danger that the need for a closer collaboration between social scientists and computer scientists which such large datasets engender may result in the treatment of the work of computer science overall as such a black box, whose in-built assumptions about the most appropriate ways to process the datasets for further analysis are never checked or discussed by the interdisciplinary team. To avoid this, it is crucial for both sides to seek to understand each other's ways of working, to learn each other's disciplinary languages, and to develop at least a basic literacy of each other's methods. For humanities and social science scholars seeking to work with 'big social data', this is likely to require the acquisition of functional mathematical and statistical knowledge and of some entry-level programming skills at the very least.

If such obstacles to successful collaboration on 'big social data' research projects can be addressed, a further challenge lies in the development of effective and appropriate data management and publication processes and protocols. Even where – as in the case of *Twitter* – researchers deal exclusively with ostensibly published material, it must be remembered that the individual users whose posts and profiles are included in the dataset are most likely unaware of the full extent of contemporary data analytics methods, and may understand *Twitter* to be a considerably more ephemeral communicative medium than it is from the researcher's or analyst's perspective. There is significant potential for social media research to engage in an in-depth profiling of individual users, but such research activities are in many cases likely to be inappropriate, especially where their outcomes are published in an identifiable or re-identifiable form; researchers will need to decide carefully, on an individual basis, what level of individual profiling is appropriate in each case. While it may be acceptable to study in detail the posting patterns of an individual, but ultimately institutional *Twitter* account such as @barackobama, for instance, the same is not true for the account of an ordinary *Twitter* user commenting on Barack Obama's policies; however, it *may* be defensible again for the account of one of the most active ordinary users @mentioning or retweeting @barackobama, as its activities will already be highly visible to other *Twitter* users. Other than to generally err on the side of caution in each case, there is no generalisable advice that can be offered to researchers on these points; rather, a careful consideration of the implications of the published research for the users whose activities it highlights is necessary in each case. A range of considered guides for addressing such questions – such as the Association of Internet Researchers' ethics guidelines (http://ethics.aoir.org/) – are now emerging from within the research community itself.

Considerable further work also still remains to be done in developing the appropriate formats for publishing 'big social data' research. At present, few published studies in relevant journals and similar environments include the raw data, due to the restrictions imposed by the terms and conditions of the *Twitter* API as much as to the limitations of contemporary publishing formats; they also leave relatively little space for a considered discussion of the methodological choices made in processing and visualising the data. Further, the static format for publishing graphs and tables which even online journals have inherited from the print format is often less than ideally suited to the publication of detailed analytics drawn from large-scale, live datasets; it would be more appropriate to publish such graphs as interrogatable, dynamic data dashboards that enable the readers (and before them, the reviewers) to replicate the analytical steps taken by the authors, and test a variety of other approaches to understanding the data patterns. The tools for creating and publishing such interactive dashboards are beginning to emerge, but their adoption remains unlikely as long as scholarly journals remain closely wedded to the print paradigm (also cf. Bruns, 2013).

Such *caveats* should not be seen as an indication that no valuable *Twitter* research drawing on 'big social data' has as yet been published – this is evidently untrue. However, for all of the initial enthusiasm and considerable energy invested into doing large-scale *Twitter* research, arguably the full potential of such approaches has yet to be realised by scholarly researchers; the field has a considerable part of the path towards greater maturity still ahead of it. Whether it can continue to progress further along this path, however, also depends on the further development of Twitter, Inc.'s increasingly problematic data access policies, and on the development of a better understanding of the advantages and pitfalls of 'big social data' research that draws on *Twitter* data. Paradoxically, the growing scale of *Twitter* research that has resulted from the gradual adoption of such computational research approaches has also served to highlight the many limitations and blind spots of 'big social data'-driven *Twitter* analysis more clearly.

The Precarity of 'Big Social Data' in a Proprietary Environment

Although *Twitter* has traditionally proven to be a fruitful environment for Internet studies, more recent changes to API access policies introduced by Twitter, Inc. have resulted in considerable disruptions to the research process. As noted, *Twitter*'s API has always been subject to access throttling, ostensibly in order to prevent frivolous, erroneous, or malicious API requests from absorbing excessive bandwidth, but such limits could be removed on a case-by-case basis by *Twitter*'s developer support team. The granting of such on-demand exceptions has ceased, and successive changes to API rules have considerably limited the number of API requests which could be made by any one client in a given 15-minute window; truly large-scale data gathering directly from the *Twitter* API is therefore no longer easily possible, except where researchers again restrict themselves to tracking relatively simplistic hashtag or keyword collections.

In its pursuit of additional revenue sources following its listing on the New York Stock Exchange, Twitter, Inc. has increasingly encouraged API users with more complex data needs to work with one of a number of commercial third-party data resellers, such as Gnip or Datasift, but these services are designed largely for major industrial clients and generally priced beyond the reach of publicly funded research institutions; notably, too, Twitter, Inc. bought up Gnip in 2014 and announced the termination of its data access arrangement with

Datasift in early 2015 (Lunden, 2015), resulting in a virtual monopoly for Twitter, Inc.'s subsidiary Gnip as the only fully authorised *Twitter* data reseller. The company has attempted to mollify scholarly researchers and other non-commercial data users by pointing to its widely publicised 2010 gift of the entire, continuously growing *Twitter* archive to the U.S. Library of Congress, but as of mid-2015 this archive remains inaccessible to researchers as Twitter, Inc. and the Library attempt to come to an agreement on the modalities of access to this resource; Twitter, Inc. also operated a competitive "Twitter Data Grants" process in 2014 which provided large-scale data access to the winning research projects, but with only six out of more than 1,300 applications chosen to receive such access (Kirkorian, 2014), the Data Grants amounted to little more than a data lottery – and the experiment appears not to have been repeated in subsequent years.

Such increasing data precarity is especially problematic given the growing need for 'big data' on the Twittersphere which the first generation of *Twitter* research activities has demonstrated more and more emphatically. Such early work has been exceptionally successful in demonstrating the incorporation of *Twitter* into a very broad range of communicative practices, and thus also in documenting the relevance of *Twitter* as a leading social media platform especially in the context of real-time public communication; Twitter, Inc. itself has materially benefitted from this research as it has encouraged government bodies and corporate partners to develop their *Twitter* presences, by being able to point to such independent scholarly assessments of the platform's importance and influence. However, at the same time this early work has also highlighted a significant lack of context for the individual case studies which have been conducted: for example, although research has been able to chart the growing user engagement with specific events and activities (such as the growth in user activity from one national election to the next, or from one Academy Awards broadcast to another), what is as yet entirely absent from the published record is any indication of the overall *Twitter* activity baselines that these specific events compare to.

While we may take note of the volume of tweets relating to the latest U.S. presidential election, for instance, it remains unclear what average level of ordinary daily tweeting activity (in the U.S., or globally) this compares to, and whether tweeting about the election has thus remained the practice of a self-selecting few who may be highly notable because of their elevated levels of participation, but are also highly unrepresentative of an otherwise politically apathetic majority – or whether, conversely, there was a truly demotic adoption of *Twitter* as a channel for discussing the election campaign as it unfolded. Fundamentally, in spite of a handful of updates from Twitter, Inc.'s internal developer team, serving largely as corporate PR, and a small number of isolated research projects such as the Silicon Graphics-supported "Global Twitter Heartbeat" project (Silicon Graphics, 2015), there is no independently verifiable scholarly data on global or national *Twitter* usage patterns, and up-to-date figures on user activity are limited to statements in Twitter, Inc.'s reports to shareholders and headline figures on *Twitter*'s 'About' page, which at the time of writing boasted 288 million "monthly active users" and 500 million tweets per day, globally (Twitter, Inc., 2015).

Even those Internet researchers who have by now developed the skills and frameworks for working with 'big social data' from *Twitter* and other social media platforms, and could thus shed additional light on more detailed usage patterns, are therefore ultimately limited by the severely restricted nature of access to truly large-scale datasets on *Twitter* usage. As Twitter, Inc.'s earlier, much more permissive and cooperative stance towards valuable, independent, publicly-funded research has evaporated, such researchers are now forced by the company's current policy to decide between investing considerable time to seek the elusive funds or corporate sponsorships that would enable them to buy access to the *Twitter* 'firehose' (the live feed of all tweets around the world), or exploring the very limits of what is acceptable under the terms and conditions of the standard *Twitter* API. They are prohibited by current terms and conditions even from publicly sharing their datasets with fellow researchers, a policy which is inherently in conflict with the increasing number of national research frameworks that require publicly-funded research to make its findings *and* its datasets public, and which – based on anecdotal evidence – a substantial number of scholarly *Twitter* researchers choose to ignore.

Such significant restrictions to working with 'big social data' from *Twitter*, and indeed the repeated, abrupt, and often ill-considered shifts in Twitter, Inc.'s data politics (cf. Puschmann & Burgess, 2014), ultimately position *Twitter* as a particularly precarious object of, and space for, data-driven research. Although comparatively simple “hashtag studies” of *Twitter*-based phenomena certainly remain possible, more complex, truly ‘big data’ work is becoming increasingly more difficult and potentially unsustainable, unless significant financial and institutional backing can be found by the researchers seeking to undertake it. Should the evolution of Twitter, Inc.’s corporate policies towards a heavy-handed commercialisation of data access continue on its present trajectory, it is possible that the ‘big data’ moment in *Twitter* research may conclude prematurely – and as the growing industry and scholarly focus on ‘big data’ highlights the utility of such data sources, this development may be a harbinger of trends in data access well beyond *Twitter* itself. ‘Big social data’, and ‘big data’ more generally, may well also turn out to be a synonym for ‘expensive data’; this is perhaps especially likely in the field of Internet studies, where so many of the online phenomena that researchers may wish to study are ultimately taking place on proprietary platforms whose operators would have the ability to control and monetise access to their data.

Conclusion: But Do We Need ‘Big Data’?

Working with ‘big data’ should never be an end in itself, of course; such data sources must be used to address meaningful questions that could not be addressed merely by using more conventional research approaches. A growing number of critical publications have pointed out that in the current ‘big data’ goldrush, such principles are at times ignored or forgotten; it must be acknowledged that there is for some researchers a fascination simply with big numbers in the datasets, even where an increase in quantity (for example in the number of tweets processed to obtain a certain result) does not measurably improve the quality of the research results. By extension, the same may also apply for funding bodies, research administrators, and the media: for instance, research outcomes based on a superficial analysis of tens of millions of tweets can turn out to be easier to present (or ‘sell’) as meaningful and representative than those resulting from in-depth interviews with a dozen users, even though it is possible that the small-scale, deep qualitative engagement with users has generated considerably greater insights than the large-scale, surface quantitative analysis of their social media posts. Part of the backlash against a headlong rush to ‘big data’ is certainly also driven by the perception that proponents of such ‘big data’ research are not treated with the level of critical scrutiny that their (and indeed, any) chosen research approaches should be objected to.

Articles such as boyd & Crawford’s “Critical Questions for Big Data” (2012), and a number of follow-on contributions which were sparked by their provocation, have helped significantly to foster such critical scrutiny; far from being uncritically celebrated, the idea of ‘big data’ has now been problematised, and – quite appropriately – the utilisation of ‘big data’ methods and resources must now be sufficiently justified in most research proposals and publications. This is true for research in the social sciences, at least, where the ‘big data’ debate has been conducted most critically and forcefully; in computer science, a somewhat less critical mindset may still prevail. To fully rehearse the arguments of that debate would be beyond the scope of the present chapter; rather, it is necessary here simply to reaffirm that ‘big social data’ research must always also critically review the provenance and quality of its datasets and the abilities and applicability of the methods used to process them, and that the aim in working with such datasets should never be the use of ‘big data’ in itself, but to use these datasets to address meaningful questions beyond the data.

By extension, this is also a call to avoid a simplistic juxtaposition of ‘qualitative’ and ‘quantitative’ research methods and approaches, as if these were always so easy to divide apart. Beyond the most basic descriptive ‘big data’-supported research, which is content simply to present an overview of the metrics of social media participation but fails to provide any further discussion and interpretation of those observations, the reality that is currently emerging in advanced *Twitter* research, for example, is one that draws inherently on mixed-methods approaches: here, the computational, quantitative evaluation of very large datasets may be utilised for instance to pinpoint specific subsets of the data that are then subjected to further qualitative analysis in the form of a close reading of tweets, or of in-depth interviews with key participants – or alternatively, an initial qualitative investigation of specific social media phenomena may provide the basis for the establishment of a corpus of key terms or a population of target accounts whose further social media careers are then tracked and analysed using large-scale qualitative methods. In the best of these projects, 'big data' from computational approaches and 'deep data' from more conventional sources are integrated to form hybrid data structures that can provide considerably more valuable insights than their constituent parts are able to do on their own.

Such research, then, is data-driven not in the negative sense that the investigation it conducts is pre-determined by the externally imposed limitations of the datasets it is able to access; rather, it is data-driven in the positive sense that a deep but explorative engagement with the newly available sources of 'big social data' enables researchers to preliminarily identify previously unknown or merely suspected patterns in communicative behaviour which can then be subjected to further rigorous analysis using mixed methods than combine qualitative and quantitative approaches. This can be described as a neither purely inductive nor entirely deductive process which instead takes an abductive approach: initially, "no logical or empirical connection is required, merely spotting patterns in the data. The results of abduction, however, are not necessarily logically or scientifically coherent; they need to be properly tested, either deductively or inductively, or both". What results is a "three-step process of abductive hypothesis forming, deductive theory construction, and inductive empirical testing" (Dixon, 2012: 201f.).

To proceed with this abductive model to developing and testing research hypotheses, 'big social data' can be extremely valuable. The approach suggested by this model is one of oversampling: gather substantially more data than would be needed to address a pre-existing research question; detect patterns in the data; then develop and test the theoretical framework that may explain such patterns. By contrast, for reasons of economy (in terms of money, time, and intellectual effort), the more conventional approach in Internet studies and related disciplines has been to gather just enough data to test and prove (or disprove) a theory. Only the recent increase in the availability of 'big social data', and of the computational tools to process them, to a wider range of researchers has made the widespread adoption of such abductive approaches possible. The large, comprehensive datasets on communicative exchanges on *Twitter* that extend beyond selected hashtags, keywords, or user populations that have become available in the course of these developments, then, provide a rich resource that can be used to develop and test many old and new questions about how human (and non-human) actors communicate at scale and in real time through contemporary online media platforms – and the work done on these questions in *Twitter* is only one example of a broader range of studies that also examine many other social and other media spaces. But crucially, such work must show an awareness of the specific implications of the methodological choices it makes – more so, arguably, than it has to date.

Further, and of most immediate importance, in the case of *Twitter* the very potential for an abductive approach to developing and testing theory which 'big social data' holds, and which has been demonstrated already at least in a handful of major studies that draw on *Twitter* data, is now intensely threatened by the gradual clamp-down on data access via the *Twitter* API in favour of commercial solutions. *Twitter* research has been a key example for the possibilities of 'big social data' research in recent years – but as scholarly researchers are locked out of higher-volume API access and priced out of the commercial data market, there is a significant danger that what remains of this field is once again only a collection of considerably more limited hashtag studies.

References

Acar, A., & Y. Muraki. (2011). Twitter for crisis communication: Lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, 7(3), 392-402.

Andrejevic, M. (2014). Surveillance in the big data era. In K.D. Pimple, ed., *Emerging Pervasive Information and Communication Technologies (PICT): Ethical Challenges, Opportunities and Safeguards*. Dordrecht: Springer, 55-69.

Arthur, P.L., & K. Bode, eds. (2014). *Advancing Digital Humanities: Research, Methods, Theories*. Houndmills: Palgrave Macmillan.

Berry, D. (2011). The Computational Turn: Thinking about the Digital Humanities. *Culture Machine*, 12, 1-22. Retrieved from http://www.culturemachine.net/index.php/cm/article/ view/440/470.

boyd, d., & K. Crawford. (2012). Critical Questions for Big Data. *Information, Communication & Society*, 15(5), 662-679.

Bruno, N. (2011). Tweet first, verify later? How real-time information is changing the coverage of worldwide crisis events. Oxford: Reuters Institute for the Study of Journalism, University of Oxford. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Tweet%20first%20%2C%20verify%20later%20How%20real-time%20information%20is%20changing%20the%20coverage%20of%20worldwide%20crisis%20events.pdf.

Bruns, A. (2014). Twitter in the 2013 Australian Election. Paper presented at the Australia New Zealand Workshop on Campaign Management and Political Marketing, Sydney, 17-18 July 2014.

Bruns, A. (2013). Faster than the Speed of Print: Reconciling "Big Data" Social Media Analysis and Academic Scholarship. *First Monday*, 18(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4879.

Bruns, A., & J. Burgess. (2011). #ausvotes: How Twitter covered the 2010 Australian federal election. *Communication, Politics & Culture*, 44(2), 37–56. Retrieved from http://eprints.qut.edu.au/47816/.

Bruns, A., & T. Highfield. (forthcoming). May the Best Tweeter Win: The Twitter Strategies of Key Campaign Accounts in the 2012 US Election. In C. Bieber & K. Kamps, eds., *The United States Presidential Election 2012*. Wiesbaden: Springer VS.

Bruns, A. & T. Sauter. (2015). Anatomie eines Trending Topics: Retweet-Ketten als Verbreitungsmechanismus für aktuelle Ereignisse. In A. Maireder, J. Ausserhofer, C. Schumann, & M. Taddicken, eds., *Tagungsband Digital Methods*. Vienna: Digital Communications Research.

Bruns, A., & S. Stieglitz. (2012). Quantitative Approaches to Comparing Communication Patterns on Twitter. *Journal of Technology in Human Services*, 30(3-4), 160-185. doi:10.1080/15228835.2012.744249.

Bruns, A., & S. Stieglitz. (2013). Towards More Systematic Twitter Analysis: Metrics for Tweeting Activities. *International Journal of Social Research Methodology*, 16(2), 91-108. doi:10.1080/13645579.2012.756095.

Bruns, A., T. Highfield, & S. Harrington. (2013). Sharing the News: Dissemination of Links to Australian News Sites on Twitter. In J. Gordon, P. Rowinski, & G. Stewart, eds., *Br(e)aking the News: Journalism, Politics and New Media*. New York: Peter Lang, 181-210.

Bruns, A., D. Woodford, & T. Sadkowsky. (2014a). Exploring the Global Demographics of Twitter. Paper presented at the Association of Internet Researchers conference, Daegu, 22-25 Oct. 2014.

Bruns, A., D. Woodford, T. Highfield, & K. Prowd. (2014b). Mapping Social TV Audiences: The Footprints of Leading Shows in the Australian Twittersphere. Paper presented at the Association of Internet Researchers conference, Daegu, 22-25 Oct. 2014.

Burgess, J., & A. Bruns. (2012). Twitter Archives and the Challenges of "Big Social Data" for Media and Communication Research. *M/C Journal*, 15(5). Retrieved from http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/561/0.

Burgess, J., & A. Bruns. (2015). Easy data, hard data: The politics and pragmatics of Twitter research after the computational turn. In G. Langlois, J. Redden, & G. Elmer, eds., *Compromised Data: From Social Media to Big Data*, London: Bloomsbury, 68-88.

Conway, B.A., K. Kenski, & D. Wang. (2013). Twitter use by presidential primary candidates during the 2012 campaign. *American Behavioral Scientist*, 57(11), 1596-1610.

Dixon, D. (2012). Analysis tool or research methodology: Is there an epistemology for patterns? In D.M. Berry, ed., *Understanding Digital Humanities*. Houndmills: Palgrave Macmillan, 191-209.

Doan, S., B.K.H. Vo, & N. Collier. (2012). An analysis of Twitter messages in the 2011 Tohoku Earthquake. In P. Kostkova, M. Szomsor, & D. Fowler, eds., *eHealth 2011*. Berlin: Springer, 58-66.

Gaffney, D. (2010). #iranElection: Quantifying online activism. *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, 26–27 Apr. 2010, Raleigh, NC*. Retrieved from http://journal.webscience.org/295/.

Golbeck, J., J.M. Grimes, & A. Rogers. (2010). Twitter use by the US Congress. *Journal of the American Society for Information Science and Technology*, 61(8), 1612-1621.

Harrington, S., T. Highfield, & A. Bruns. (2013). More than a Backchannel: Twitter and Television. *Participations: Journal of Audience & Reception Studies*, 10(1), 405–409. Retrieved from http://www.participations.org/Volume 10/Issue 1/30 Harrington et al 10.1.pdf.

Highfield, T., S. Harrington, & A. Bruns. (2013). Twitter as a Technology for Audiencing and Fandom: The #Eurovision Phenomenon. *Information, Communication & Society*, 16(3), 315–39. doi:10.1080/1369118X.2012.756053.

Hughes, A.L., & L. Palen. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3-4): 248-260.

Hughes, A.L., L.A. St Denis, L. Palen, & K.M. Anderson. (2014). Online public communications by police & fire services during the 2012 Hurricane Sandy. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1505-1514.

Kirkorian, R. (2014). Twitter #DataGrants Selections. *Twitter Engineering Blog* 17 Apr. 2014. Retrieved from https://blog.twitter.com/2014/twitter-datagrants-selections.

Larsson, A.O., & H. Moe. (2011). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, 14(5), 729-747. doi:10.1177/1461444811422894.

Lotan, G., E. Graeff, M. Ananny, D. Gaffney, I. Pearce, & d. boyd. (2011). The Arab Spring: The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, 5, 1375-1405.

Lunden, I. (2015). Twitter Cuts Off DataSift to Step Up Its Own Big Data Business. *Techcrunch* 11 Apr. 2015. Retrieved from http://techcrunch.com/2015/04/11/twitter-cuts-off-datasift-to-step-up-its-own-b2b-big-data-analytics-business.

Manovich, L. (2012). Trending: The Promises and the Challenges of Big Social Data. In M.K. Gold, ed., *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, 460-475.

Mendoza, M., B. Poblete, & C. Castillo. (2010). Twitter under crisis: can we trust what we RT? Paper presented at the 1st Workshop on Social Media Analytics (SOMA '10). Washington, DC: ACM.

Meraz, S., & Z. Papacharissi. (2013). Networked gatekeeping and networked framing on #egypt. *International Journal of the Press and Politics*, 18(2), 138-166. doi:10.1177/194016121247447.

Mourtada, R., & F. Salem. (2011). Civil movements: The impact of Facebook and Twitter. *Arab Social Media Report*, 1(2). Retrieved from http://www.dsg.ae/En/Publication/Pdf_En/DSG_Arab_Social_Media_Report_No_2.pdf.

Palen, L., K. Starbird, S. Vieweg, & A. Hughes. (2010). 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.

Puschmann, C., & J. Burgess. (2014). The Politics of Twitter Data. In K. Weller et al., eds., *Twitter and Society*. New York: Peter Lang, 43-54.

Rogers, R. (2009). *The End of the Virtual: Digital Methods*. Amsterdam: Vossiuspers UvA. Retrieved from http://www.govcom.org/publications/full_list/oratie_Rogers_2009_preprint.pdf.

Rogers, R., F. Jansen, M. Stevenson, & E. Weltevrede. (2009). Mapping democracy. Paper presented at Global Information Society Watch 2009, Association for Progressive Communications and Hivos. Retrieved from http://www.giswatch.org/sites/default/files/mappingdemocracy.pdf.

Sarcevic, A., L. Palen, J. White, K. Starbird, M. Bagdouri, & K. Anderson. (2012). 'Beacons of hope' in decentralized coordination: Learning from on-the-ground medical twitterers during the 2010 Haiti earthquake. Retrieved from http://www.cs.colorado.edu/~palen/Home/Crisis_Informatics_files/Sarcevic-et-al-HaitiMedicalTwitterers.pdf.

Silicon Graphics. (2015). *Global Twitter Heartbeat*. Retrieved from http://www.sgi.com/go/twitter/.

Twitter, Inc. (2015). About Twitter, Inc. Retrieved from https://about.twitter.com/company.

Vergeer, M., & L. Hermans. (2013). Campaigning on Twitter: Microblogging and Online Social Networking as Campaign Tools in the 2010 General Elections in the Netherlands. *Journal of Computer-Mediated Communication*, 18(4), 399-419.