Methodological Innovation in Precarious Spaces: The Case of Twitter

Axel Bruns and Jean Burgess

ARC Centre of Excellence for Creative Industries and Innovation

Queensland University of Technology

Brisbane, Australia

a.bruns / je.burgess @ qut.edu.au

Keywords: Twitter, big data, social media analytics, data politics, data access, Application Programming Interface, research methods

Introduction: Social Media Analytics and the Politics of Data Access

The contemporary social media moment can be understood in terms of a 'platform paradigm' (Burgess, 2014) – one in which the private, interpersonal and public communication of a significant majority of users is being mediated via a small number of large proprietary platforms like *Facebook* and *Twitter,* and those platforms are redefining how such communication can be monetised and analysed. In this current conjuncture the data generated either directly or indirectly by user practices and interactions are at the centre of such platforms' business models – user data analytics are used to power advertising and personalise newsfeeds; and user-created social media content is in itself a commodity to be mined commercially for business insights, PR crisis aversion and even stock market prediction. Alongside such commercially motivated developments, the social and behavioural sciences as well as the digital humanities have been developing ever more sophisticated and large-scale methods for analysing social media data, often motivated by different questions but relying on similar tools to access and analyse data as the commercial

players, and thereby operating in ways that entangle scientific practice with the evolving markets in user data. To complicate matters, as the power and uses of social data analytics have grown, so too has the social anxiety around surveillance, exploitation and user agency.

While such multiple interests intersect, compete, and conflict in and around the issue of access to and use of social media data (Puschmann & Burgess, 2014), here we are most interested in those uses which are explicitly framed in terms of *research*, and therefore for the purposes of clarity in this chapter, we concentrate on key differences between commercial, market-oriented research and scholarly, scientific research. Commercial research is frequently centred around three main themes: approaches which enable platform providers themselves to better understand their users and ensure that further technological enhancements meet their needs and interests; approaches which allow the advertisers and marketers that contribute to the platforms' revenues to more effectively target specific interest groups within the overall userbase; and approaches which enable corporate players and other institutional actors to understand and improve the ways their customers are engaging with them as a brand or as a company. Scientific research using social media data has expanded beyond the early interests of computer and information scientists on the one hand and pockets of the humanities and social sciences on the other to include a wide range of social, behavioural, and even physical science disciplines interested in how 'naturally' occurring social interaction data can be mined to understand the dynamics of self-organising systems, information diffusion, and social influence. In the field of communication, large-scale, data-driven social media research tends to be motivated by questions about the systemic communicative processes which are evident within a large and diverse user population, and on

investigating how such processes respond to specific short-term events within and beyond the social media platform itself. There are also considerable points of connection between scientific and commercial research interests, of course, and in spite of potentially substantial differences in the ethical and organisational frameworks which govern their research and the very real conflicts that these differences can produce – as the Facebook 'emotional contagion' controversy demonstrates (Kramer *et al*., 2014) – fruitful collaborations are possible.

Regardless of the commercial or scientific orientation of individual research projects, the fundamental point must also be made that social media research as such is genuinely important, for a variety of reasons. Social media have now become a major form of public communication in their own right. Indeed, they are one of the few truly *public* forms of communication currently available, in the sense that they enable billions around the world to publicly express their thoughts, ideas, interests, ambitions, likes and dislikes within a shared global communications environment. This does not mean that all such voices are equally audible, of course, but it is precisely the dynamics of how specific issues, themes, and memes emerge to prominence from this globally distributed conversation, and what impact they may come to have well beyond individual social media platforms themselves, that has become a key object of study for social media researchers across fields from political through crisis to enthusiast and everyday communication.

Increasingly central to both the commercial and scientific research agendas, therefore, has been the development of social media analytics methodologies which are able to draw on large and potentially real-time datasets that describe the activities (or at least those activities which are publicly visible to other participants) of a large number of social media users. The current generation of social media

platforms is distinct from its predecessors in part due to its greater focus on the multi-platform use and embeddability of its content, enabling users to use a range of official and third-party tools to access their social media feeds across different devices and operating systems as well as allowing various parties to embed relevant social media feeds and functionality within websites, smartphone and tablet applications (apps), and other contexts. Such functionality is supported by modern social media platforms chiefly through the provision of a unified and well-documented Application Programming Interface (API): an interface which constitutes an access point that, on request, provides structured data in a standard format which does not prescribe the context or form in which such data are made available to the end user. While such APIs are used mainly by popular social media end-user clients from the official *Facebook* and *Twitter* apps to *Tweetdeck* and *Hootsuite*, they also provide an exceptionally useful point of access to social media data for researchers. Using APIs it becomes possible to retrieve the public profile information and public updates posted by specific users or containing given keywords or hashtags, for example; processed effectively, such data become the raw material for social media analytics.

At the most basic level, analytics approaches which draw on the APIs provided by leading social media platforms are necessarily limited by the range and amount of data available through Application Programming Interfaces. Application Programming Interfaces rarely provide unrestricted access to the totality of all data about users and their activities that may be available internally; for example, data about private (i.e. non-public) messages exchanged between individual users are available generally only to these users themselves, and to the API clients to which they have provided their authorisation codes. Such restrictions result in considerable

differences in what social media analytics approaches are able to investigate for different social media platforms, then: on *Facebook*, for example, few posts (except for posts and comments on public pages, and posts on user profiles whose visibility level has been explicitly set to 'public') are truly globally public, while a majority is visible only to the sender's 'friends' or 'friends of friends'. Unless it has been authenticated by a user within such a friendship circle, such semi-private posts will remain invisible to a tool gathering social media data. *Twitter*, on the other hand, uses considerably more limited privacy settings: its accounts are either 'public' (meaning that all of their tweets are globally public, and visible even to non-registered visitors to *twitter.com*) or 'protected' (tweets are visible only to followers of the account whom the user has explicitly approved). Since only a small and shrinking minority of *Twitter* accounts are set to be 'protected' in this way, the activity data potentially available through the *Twitter* API therefore constitutes a considerably more comprehensive reflection of the totality of *Twitter* activity than is the case for *Facebook*, where a dataset of globally public posts would contain only an unpredictable (and certainly unrepresentative) mixture of deliberately and accidentally 'public' messages.

In spite of *Twitter*'s smaller global user base – as of July 2014, it claimed 271 million 'monthly active users' (Twitter, Inc., 2014), compared to *Facebook*'s 1.23 billion (PR Newswire, 2014) –, social media analytics methodologies for *Twitter*, especially where they draw on large datasets tracking the activities of users, are therefore arguably more developed than those for *Facebook* and have begun to generate new and important insights not only into how *Twitter* itself functions as a social media platform, but also into how the patterns of user activity found here can

be seen as exemplary for the adoption, adaptation, and use of new, digital communications technologies more generally.

But *Twitter*'s APIs are far from neutral and transparent tools. Rather, APIs are an essential means for the platform provider to encourage some uses of user data and to regulate or even prohibit others – affecting the research agenda and business plans of all those who would make use of user data. Secondly, in addition to constraining and enabling particular *kinds* of data use, the *Twitter* APIs have changed over time as the business imperatives of the platform have changed, often in ways that are misaligned with third-party developers and other actors in the *Twitter* 'ecosystem'. *Twitter*'s APIs therefore mediate between and are the site of friction between competing uses and understandings of Twitter as a platform, and changes to how they work have been accompanied by controversy and debate – as Taina Bucher has argued in work that reports on interviews with third-party developers working with *Twitter* data, APIs are 'objects of intense feeling' (Bucher, 2013, n.p.). As data-driven *Twitter* research began to grow in scope and in the stakes attached to it, such shifts have also become increasingly politicised and materially significant for the scientific community. Academic researchers have been no less frustrated and entangled with the politics of these APIs, which sit alongside other practical and ethical challenges in doing data-driven social media research (see Lomborg & Bechmann, 2014, for an excellent overview).

Consequently, this chapter focusses substantively on the changing affordances of *Twitter* data, as well as the tools and methods for analysing it, with reference to questions of methodological advancement in our core disciplines of journalism, media, communication, and cultural studies. But at the same time, this story can reveal as much about the political economy of the new digital media

environment as it does about the pragmatics of scientific research in this environment. This chapter contributes to such an understanding via a short history of the uses of *Twitter* data for media, communication and cultural research, the methodological innovation that has taken place over this time, and the stakeholder relationships and sociotechnical arrangements that have both supported and constrained such work.

Phase 1: Building the *Twitter* Ecosystem

Although some early *Twitter* research drew on more primitive methods for gathering data from the platform – such as taking regular screenshots or using generic HTML scrapers to regularly archive the *Twitter* feeds of selected users – the considerably greater utility of instead connecting to the API to gather data in a standardised and reliable format soon led researchers to pursue that avenue. At first, the tools used to gather data from the API were mainly developed *ad hoc* and in house at various research institutions; for the most part, they focussed initially on gathering the tweets posted by selected accounts, or containing specified keywords or hashtags. (Our own contributions to this effort, building on open-source technologies, are gathered in Bruns & Burgess, 2011c, for example.)

The *Twitter* API imposes a number of restrictions on its users, relating to the number of users and search terms which may be tracked through a single API request, and to the volume of data which is returned. At the time of writing, for example, the open API only returns up to one per cent of the total current volume of tweets being posted. This means that if, this hour, *Twitter* users around the world were posting one million tweets in total, a keyword search for 'twitter' would return only up to 10,000 tweets during that hour, even if more tweets had contained the

term ‘twitter’. The API will also notify its client about how many tweets had been missed, however. In a variety of contexts, such restrictions pose significant complications: research which tracks common keywords such as ‘flood’ or ‘earthquake’ to extract early indicators of impending natural disasters would be severely limited by the throttling of its data access at one per cent, for example, especially at times when one or more severe disasters coincide. However, current literature which studies the uses of social media in crisis communication by drawing on *Twitter* datasets largely omits any discussion of this potentially crucial limitation.

Both to address such issues and to more generally encourage the development of innovative *Twitter* analytics models, Twitter, Inc. therefore instituted an API whitelisting system for developers and researchers. By contacting Twitter support staff, interested third parties could request a lifting of API access restrictions for the data gathering tools they developed. With whitelisted access made available *ad hoc* and relatively speedily, this supported the emergence of a number of popular *Twitter* clients for professional end-users (such as *TweetDeck or Hootsuite*), as well as the development of a range of research initiatives which aimed to work with larger *Twitter* datasets than were commonly available to API users. Twitter, Inc. itself recognised the importance of this growing “ecosystem” of developers and tools which drew on and enhanced the central platform; indeed, the research outcomes enabled by this early, explorative phase of *Twitter* analytics also contributed substantially to demonstrating the importance of *Twitter* as a platform for public communication in contexts ranging from second-screen television viewing (see e.g. Highfield *et al*. on *Twitter* and the Eurovision Song Contest, 2013) to political activism (see e.g. Papacharissi & de Fatima Oliviera on *Twitter* in the Egyptian

uprising, 2013), and served to establish *Twitter* as one of the most important global social media platforms.

Finally, this early phase of research innovation also resulted in a first trend towards methodological consolidation, as several leading tools for gathering *Twitter* data emerged. These included stand-alone tools such as *140kit* and *TwapperKeeper* as well as the *Google Spreadsheets* extension *TAGS* (cf. Gaffney & Puschmann, 2014). The growing use of such publicly available tools in preference to in-house solutions meant that the datasets gathered by different researchers and teams were now more immediately comparable, and enabled the development of a range of standard analytical tools and metrics building on common data formats (Bruns & Stieglitz, 2013). This also considerably enhanced the level of scholarly rigour in social media analytics by enabling researchers to replicate and test each other's methodological frameworks. The availability of such tools as free hosted services, or as software released under open source licences, also contributed significantly to such methodological innovation and evaluation: the open availability and extensibility of the key early research tools instilled a strong 'open science' ethos in the international *Twitter* research community which gathered around these tools and methods.

The common focus of many of these emerging tools on enabling, in the first place, the tracking of set keywords and – especially – hashtags also resulted in the emergence of an increasingly dominant subset of *Twitter* analytics which is best summarised under the title of 'hashtag studies': research initiatives which sought to capture a comprehensive set of tweets containing prominent hashtags relating to specific themes and events, from natural disasters – e.g. #terremotochile (Mendoza *et al*., 2010) – to national elections – e.g. #ausvotes (Bruns & Burgess, 2011a). Such

hashtag studies built on the tendency of *Twitter* users to self-select some of their tweets as relevant to specific topics by including a topical hashtag in the tweet text, and generated considerable new insights into the self-organising nature of *ad hoc* communities on *Twitter* (Bruns & Burgess, 2011b). However, they also captured only a very specific range of user practices taking place especially around acute events, while being unable to meaningfully investigate the arguably more commonplace practices of non-hashtagged everyday and phatic communication on *Twitter*. Following the distinctions proposed in Bruns & Moe (2014), such hashtag studies focus largely on the macro-level of *Twitter* communication which builds on hashtags, while ignoring the meso-level (everyday interaction with one's followers) and the micro-level (public conversations using @replies).

The early popularity of hashtag studies also resulted from the fact that the availability of tools such as *TwapperKeeper* in the form of a Web-based service enabled even researchers with minimal technical skills to track and gather sizeable datasets of all tweets containing specified hashtags (and keywords), limited only by the API restrictions imposed by Twitter, Inc. *TwapperKeeper* simply required researchers to enter their keywords, and provided a Web interface to download the resultant datasets in user-friendly Excel or comma-separated values formats. Further, such datasets – once gathered – were made available to all users of the site, and *TwapperKeeper.com* therefore became a *de facto* clearinghouse for *Twitter* archives.

However, as *TwapperKeeper*'s popularity and use increased, Twitter, Inc. gradually developed the view that its Web-based service – and especially its provision of public archives of hashtag and keyword datasets – contravened the Terms of Service of the *Twitter* API, which prohibited the public sharing of API-

derived data in programmatic form. In early 2011, Twitter, Inc. ordered *TwapperKeeper.com* to cease its public service (O'Brien, 2011); subsequently, some of *TwapperKeeper*'s archival functionality was incorporated into third-party client application *Hootsuite*, while the source code for a DIY version of *TwapperKeeper*, *yourTwapperKeeper*, was made available publicly by developer John O'Brien III under an open source licence. Arguably, this moment is emblematic for a more fundamental and significant shift in Twitter, Inc.'s relationship with the developer and researcher community and ecosystem which had developed around its platform, and marks the beginning of a second, considerably more precarious phase for innovation in *Twitter* analytics.

Phase 2: Precarious Access as Demand for 'Big Data' Grows

Twitter, Inc.'s increasingly restrictive interpretation of the APIs Terms of Service, its attendant discontinuation of whitelisting practices, and overall changes to API functionality and access limitations since 2011 constituted a disruption of the established equilibrium between platform provider and third-party developers and researchers that, while undermining many existing research methods and approaches for *Twitter* analytics, also resulted in considerable new innovation and development. During the first phase of methodological innovation around *Twitter*, developers and researchers had relied at least implicitly on Twitter, Inc.'s continued good will and support towards them, and even – as in the case of *TwapperKeeper* and the wider whitelisting regime – on a willingness on part of the platform provider to bend its own rules and overlook what could be considered to be breaches of its API rules. As Twitter, Inc. began to withdraw its support for the third-party ecosystem which had played a substantial role in bringing *Twitter* to a position of global

prominence, developers scrambled to revise their methods and tools in a way that would not – or at least not obviously – put them in conflict with the company's new, stricter rules, or would devolve responsibility for any transgressions from the developers to the end-users of their data gathering and processing tools.

This shift can be observed in the transition from *TwapperKeeper* to *yourTwapperKeeper*. While the former offered data gathering functionality as a Web-based service, the latter simply provided an open-source version of *TwapperKeeper* functionality as a package which interested and sufficiently skilled researchers could install on their own servers, and could use and even modify as required for their specific purposes. Unless steps are taken to specifically prevent such access, *yourTwapperKeeper* installations continue to make their archives of gathered data available for download to anybody – not just to the researchers who entered the search terms to be tracked. This breach of the API's Terms of Service is a matter for the administrators of each individual *yourTwapperKeeper* server instance, not for *yTK*'s developers, and Twitter, Inc. would need to pursue these administrators individually if it aimed to comprehensively shut down any unauthorised sharing of API-derived *Twitter* datasets; to date, it has not attempted to do so. At the same time, the *TwapperKeeper* experience and the implicit threat of cease-and-desist requests from Twitter, Inc. have generally led researchers and institutions operating *yourTwapperKeeper* instances and similar tools to refrain from publicly advertising such services and sharing their datasets: Twitter, Inc.'s very public shutdown of *TwapperKeeper.com* in March 2011 can be said to have had a notable chilling effect on the sharing of data in the international social media researchers' community.

Conversely, the *TwapperKeeper* shutdown has led that community to increase its efforts to develop better tools for tracking, gathering, processing, and analysing

social media data at large scale. In addition to *yourTwapperKeeper* and its derivatives, such tools also include projects such as the *Twitter Capture and Analysis Toolset* (*TCAT*), developed by the University of Amsterdam's Digital Methods Initiative (DMI, 2014), which similarly requires users to install their own *TCAT* instance on a server they administer; advancing beyond the mere tweet archiving functionality provided by *yTK*, *TCAT* also offers a range of built-in analytics functions which provide first quality control and quantitative insights into the data being gathered. Such new advances in the development of more powerful and complex yet still Terms of Service-compatible *Twitter* research tools also create new divides within the established social media researchers' community, however. They separate researchers and teams who possess the necessary technical expertise to install and operate server-side solutions for data gathering and analysis (now crucially including computer science and related skills) from those who were able to work with the datasets provided by the previous generation of Web-based data gathering services but find themselves unable to operate their own servers. As the capabilities, but also the complexity of server-side tools grow, this presents a very tangible risk of dividing researchers into 'big data' haves and have-nots.

Such divisions are also emerging, on a much larger scale, between unfunded and publicly funded scientific research initiatives using open-source tools connecting to the standard *Twitter* API on the one hand, and commercial research projects and companies buying social media data at more substantial volumes from third-party suppliers on the other. Twitter, Inc.'s agenda in tightening open access restrictions to the public API from 2011 onwards was evidently also aiming to push those API clients who could afford it to make use of available third-party services such as Gnip and DataSift, which had been accredited by Twitter, Inc. as commercial data

resellers. (Gnip itself has since become a wholly-owned subsidiary of Twitter, Inc. itself.) Using such services, it is possible to buy access to tweets in high-volume keyword feeds or from large user populations, or even to comprehensive global *Twitter* feeds up to and including the full 'firehose' of all tweets, without the limitations in the depth or speed of access imposed by the public API – however, this will commonly generate costs in the tens of thousands of dollars for large one-off data purchases, and even higher cumulative costs for longer-term data subscriptions. Additionally, DataSift provides access to historical data, which is not available from the API. The volume prices quoted by resellers such as Gnip and DataSift render such services unaffordable for researchers without considerable corporate, institutional, or grant funding, however; to date, only a small number of scientific research initiatives are known to have bought data from these providers, which otherwise mainly service commercial market research services. The vast majority of researchers at public research institutions continue to draw on the public API service, and thus remain at the mercy of Twitter, Inc.'s decisions about API functionality, access limitations, and Terms of Service.

Several statements by Twitter, Inc. that acknowledge the importance of *Twitter* data as an unprecedented record of public communication activities, and of independent scientific research as shedding new light on the user practices contained in such data, may be seen as seeking to address this troubling divide between data-rich commercial marketing research and data-poor publicly-funded research. In 2010, the company gifted a complete and continuously updated archive of all tweets ever sent to the U.S. Library of Congress, which the Library has subsequently sought to make available to selected researchers. In 2013, it instituted a competition for "Twitter Data Grants" which are set to provide direct access to

Twitter data at high volume. However, neither of these initiatives have so far been able to meaningfully address the lack of affordable large-scale access to *Twitter* data for publicly-funded scientific research. Access to the Library of Congress's comprehensive dataset has been stalled both by the technical challenges of making searchable an archive of billions of individual messages, and by difficult negotiations with Twitter, Inc. over the conditions of access to the archive, and only in 2013 has the Library finally offered access to its *Twitter* archive to the three winners of its annual Kluge Fellowship in Digital Studies (Library of Congress, n.d.). Similarly, in 2014 Twitter, Inc. selected only six winners from more than 1,300 applicants in the inaugural round of its Data Grants competition (Kirkorian, 2014). Even taken together, these nine grants cannot but fail to address the lack of access to 'big data' on *Twitter* activities now experienced by scientific social media research.

It must be noted at this point that scientific research into social media uses and practices is not always automatically enhanced and improved by access to larger datasets; as boyd and Crawford (2012) have shown, 'big data' does not always mean 'better data', and important social media research is being done by using comparatively small but rich datasets on social media activities which were gathered through means other than by requesting data from the APIs of *Twitter* itself or of third-party data resellers. However, for the purposes of this article we are concerned specifically with social media *analytics* as a subset of a wider and more diverse range of social media research methodologies, and this area of social media research is defined largely by its predominantly quantitative approach to working with social media data. Such quantitative analytics also remain possible for smaller datasets, of course – but to put even such analyses of smaller datasets into context (for example, to benchmark *Twitter* activities around acute events against longer-

term baselines), ‘big data’ on social media usage patterns across larger user populations and long-term timeframes are indispensable. The development of social media analytics as a serious scientific field crucially depends on researchers’ access to ‘big data’ on the use of social media platforms such as *Twitter*,.

Phase 3: Crash or Crash Through?

In the absence of affordable, or even of available options for accessing ‘big data’ on public communication using social media platforms such as *Twitter*, there is anecdotal evidence that a growing number of researchers are prepared to explore the very limits of the *Twitter* API, and in doing so also of Twitter, Inc.’s interest in strictly enforcing its Terms of Service. We have already seen that even during the earlier, comparatively permissive phases of the development of social media analytics using *Twitter* data, researchers were frequently sharing their datasets with each other – even if to do so was likely to constitute a breach of the Terms of Service under which API data were provided. In this context, Twitter, Inc.’s rules for data provision come into direct conflict with standard scientific practice: first, the open publication of raw datasets is generally encouraged as such datasets are often indispensable for an independent verification of a researcher’s findings by their peers; second, public funding bodies such as the Australian Research Council or the U.K. Arts and Humanities Research Council are increasingly requiring the data and results generated by the projects they fund to be made available publicly under open access models. While exceptions to such rules are commonly made for datasets which are commercial in confidence or otherwise restricted from publication, an argument for such restrictions is difficult to sustain in the case of *Twitter* datasets retrieved from a public Application Programming Interface and containing public

messages which – by Twitter, Inc.'s own Terms of Service (Twitter, Inc., 2012) remain copyrighted to their original senders. At least in principle then, the further distribution amongst researchers of datasets containing tweets should put those researches in potential conflict mainly with those *Twitter* users, not with Twitter, Inc.

While such arguments, as well as the overall applicability and force of *Twitter*'s Terms of Service (for both *Twitter* overall, and for the API in particular) in relation to user and researcher rights and obligations, has yet to be tested in full and across various national jurisdictions, it is therefore at least understandable that many researchers appear prepared to bend the API Terms of Service by sharing datasets at least privately, in order to meet their obligations to their scientific peers and public funding bodies. Especially for *Twitter* researchers working in project teams (for example in the context of formal, funded research projects) rather than as sole operators, such sharing is ultimately inevitable, as they must necessarily develop a shared repository of the data gathered in pursuit of the team's research agenda. Even such intra-team sharing – for example by establishing a *yourTwapperKeeper* or *TCAT* server utilised by members of the research team – may already be seen as contravening the API Terms of Service's prohibitions against "exporting Twitter Content to a datastore as a service or other cloud based service" (Twitter, Inc., 2013).

It is unlikely that Twitter, Inc. would seek to enforce such a narrow interpretation of its rules, but this in turn creates further confusion for researchers. If intra-team sharing of datasets is permissible at least implicitly, then – given the vagaries of what constitutes a research team – where are the limits to such sharing? If, for instance, a small project team funded for a brief period of time is allowed to operate a *TCAT* server and share its datasets amongst the team members, could that permission be

extended to the members of a larger, indefinitely continuing research group, centre, or institute, or even to an entire university? If multiple universities formed a consortium collaborating on joint social media research projects, could their datasets be shared across all member institutions? In the absence of clear guidance from Twitter, Inc. on such matters, as well as of independent legal advice on the validity of such guidance within their home jurisdiction, it is likely that many researchers will continue to be prepared to push the envelope further, at least until Twitter, Inc. reprimands them for their actions.

Similar “crash or crash through” approaches may emerge at a more purely technical level. At present, *Twitter*’s public API is throttled in a number of aspects, as we have already noted. In addition to the fundamental restriction that no client connecting to the streaming API (which provides real-time *Twitter* activity data) is able to retrieve more than one per cent of the total current volume of *Twitter* activity, other API calls (for example to the search API, which delivers recent tweets matching specific criteria, or to the user API, which provides information on public user profiles) are throttled by accepting only a limited number of calls from the same client in each 15-minute time window, as well as by delivering large results lists in a paged format that requires multiple API calls. Such limits do not entirely disable, but certainly significantly slow the retrieval of large datasets through such API calls – and it is again likely that such throttling is designed to promote the use of commercial data reselling services instead of the public API.

Provided that sufficient development expertise is available, it is obvious that such per-client access limits can be circumvented comparatively easily by substantially parallelising API calls. Under this model, as soon as one API client reaches the access limit for the current 15-minute window, another takes over until the next

window begins. Here, too, it appears that the extent to which such parallelisation is in explicit breach of the API Terms of Service has yet to be tested, especially as few researchers exploring such approaches are likely to publicly advertise this fact. Twitter, Inc.'s adjustments to and variable enforcement of its Terms of Service over recent years have created substantial levels of mistrust between the company itself and the social media research community that investigates how its platform is being used for public communication. This has resulted in a chilling effect which has led some cutting-edge methodological innovation to operate with considerable secrecy and under precarious conditions. This perceived need to operate more clandestinely has also severely undermined the earlier 'open science' ethos of the *Twitter* research community, of course – detailed discussions of such advanced methods are unlikely to take place in public now, for fear of reprisals from Twitter, Inc.

Conclusion: Beyond Precarity

The current trajectory of social media analytics – and of *Twitter* analytics in particular –, as we have described it here, is largely untenable. Twitter, Inc.'s interventions in the developer ecosystem, made largely by adjusting its API Terms of Service and their enforcement, as well as by throttling the functionality of the public API, have resulted in a divide between private market research institutions able to afford the commercial data access fees charged by third-party resellers and public, scientific research initiatives forced to make do with the public *Twitter* API. Internally, this latter group is further divided according to scientific researchers' ability to use existing or develop new server-side data gathering and analysis tools, and their preparedness to bend the API rules and limitations in order to access the large datasets required to develop more comprehensive social media analytics.

Faced with such challenges, it is tempting to suggest that researchers would be better advised to divert their energy to a more fertile object of investigation than *Twitter* has now become, but – while some researchers may have indeed done so – this too is an unsatisfactory option. First, the widespread adoption of *Twitter* as a tool for public communication and debate across a range of fields (from political debate through crisis communication to everyday sociality) means that it is now an important medium whose role across these fields must be researched in detail. In the field of crisis communication alone, for example, it is crucial that researchers investigate how *Twitter* is used to disseminate information during acute events, and how emergency management organisations may engage with and enhance such processes. Second, given that importance, the conduct of such research cannot be left to commercial market research institutions alone, most of which would pursue only a very limited range of research questions that are likely to generate an immediate commercial return on investment. Rather, what is needed in addition to such instrumental and applied research is the pursuit of the much more fundamental methodological and research agendas which will ultimately come to inform such applied research.

If it is important that fundamental scientific research in the field of social media analytics be conducted, and that such research include *Twitter* as an especially important platform for public communication, the current precarity of scientific research into *Twitter* and its uses must be addressed as a matter of priority. This is likely to require several concurrent initiatives: first, researchers' home institutions and funding bodies must be prepared to redress the balance between Twitter, Inc.'s commercial agenda on the one hand, and the requirements of rigorous scientific engagement with *Twitter* as a space for public communication on the other. Where

necessary, this may have to include a testing of the applicability and legality of the API Terms of Service within relevant jurisdictions. Second, there is a need to articulate more clearly and forcefully to Twitter, Inc. the value of the scientific research into the uses of its platform which has been and is being conducted. While such research has been and must be undertaken without predetermining an outcome, it is evident that most of the findings to date have demonstrated the substantial public value of *Twitter* as a new and largely open-access medium, and such findings have contributed significantly to shifting public perceptions of the platform from being for solipsistic "what I had for lunch" statements to supporting meaningful engagement across many contexts at a personal as well as professional level: as Rogers (2014) has demonstrated, scientific research has contributed to and even substantially led the 'debanalization' of *Twitter*. Third, there is a clear and urgent need to develop transparent and mutually beneficial collaborations between scientific and commercial researchers and their home institutions in order to facilitate a continuous conversation about research methodologies, ethics, and results, to enable effective and accountable processes of researcher training and knowledge transfer, and to ensure the rigorous validation of commercially funded and supported research against scientific criteria. To date, there are a small number of corporate-hosted research labs (including Microsoft Research and Intel Labs) which conduct social media research at scientific standards, and partner with universities and other recognised scientific organisations, without pursuing an inherent corporate agenda. Such industry support for genuine scientific research must be broadened further, especially at a time of limited public funding for scholarly research.

In future, by contrast, if meaningful scientific inquiry into the uses of *Twitter* is further marginalised in favour of commercially motivated studies by Twitter, Inc.'s

policies of data access, there is a real risk that the platform may be rebanalised by commercial studies that amount to little more than counts of which celebrity has attracted the most followers or which brands have generated the greatest number of retweets. Similarly, if the capability to conduct 'big data' social media research at scientific levels of accountability and rigour is concentrated in only a handful of corporate-sponsored research labs, there is a significant danger that this concentration and contraction of scholarly social media research threatens the equity of access to research methods and limits the breadth and depth of scientific inquiry and methodological innovation in this important emerging field of research. Such developments are no more in the interests of Twitter, Inc. itself than they are in the interest of the scientific research community which has established and continues to develop the fledgling field of social media analytics. The research community itself can fight to avert such developments, but only Twitter, Inc. is able to stop them, by reconsidering the frameworks which govern how it provides large-scale data access to scientific researchers.

References

boyd, d., and K. Crawford (2012) 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon', *Information, Communication & Society*, 15(5), 662-679.

Bruns, A., and J. Burgess (2011a) '#ausvotes: How Twitter Covered the 2010 Australian Federal Election', *Communication, Politics & Culture*, 44(2), 37-56.

Bruns, A., and J. Burgess (2011b) 'The Use of Twitter Hashtags in the Formation of Ad Hoc Publics', paper presented at the European Consortium for Political Research conference, Reykjavík, 25-27 Aug. 2011. Retrieved 1 Sep. 2014 from

http://eprints.qut.edu.au/46515/.

Bruns, A., and J. Burgess. (2011c, 22 June). Gawk Scripts for Twitter Processing. *Mapping Online Publics*. Retrieved 30 Jan. 2015 from http://mappingonlinepublics.net/resources/.

Bruns, A., and H. Moe (2014) 'Structural Layers of Communication on Twitter', in K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann (eds.) *Twitter and Society* (New York: Peter Lang), 15-28.

Bruns, A., and 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.

Bucher, T. (2013) 'Objects of Intense Feeling: The Case of the Twitter APIs', *Computational Culture*, 3. Retrieved 1 Sep. 2014 from http://computationalculture.net/article/objects-of-intense-feeling-the-case-of-the-twitter-api.

Burgess, J. (2014) 'From "Broadcast Yourself" to "Follow Your Interests": Making Over Social Media', *International Journal of Cultural Studies*. Retrieved 1 Sep. 2014 from http://ics.sagepub.com/content/early/2014/01/13/1367877913513684.abstract.

Digital Methods Initiative (DMI) (2014, 12 June) 'Twitter Capture and Analysis Toolset (DMI-TCAT)'. Retrieved 1 Sep. 2014 from https://wiki.digitalmethods.net/Dmi/ToolDmiTcat.

Gaffney, D., and C. Puschmann (2014) 'Data Collection on Twitter', in K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann (eds.), *Twitter and Society* (New York: Peter Lang), 55-68.

Highfield, T., S. Harrington, and 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

Kirkorian, R. (2014, 17 Apr.) 'Twitter #DataGrants Selections', *Twitter Engineering Blog*. Retrieved 1 Sep. 2014 from https://blog.twitter.com/2014/twitter-datagrants-selections.

Kramer, A.D.I., J.E. Guillory, and J.T. Hancock (2014) 'Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks', *Proceedings of the National Academy of Sciences*, 111(24), 8788-8790. Retrieved 1 Sep. 2014 from http://www.pnas.org/content/early/2014/05/29/1320040111.full.pdf.

Library of Congress (n.d.) 'Kluge Fellowship in Digital Studies Description'. Retrieved 1 Sep. 2014 from http://www.loc.gov/loc/kluge/fellowships/kluge-digital.html?loclr=blogsig.

Lomborg, S., and A. Bechmann (2014) 'Using APIs for Data Collection on Social Media', *The Information Society*, 34(4). DOI: 10.1080/01972243.2014.915276.

Mendoza, M., B. Poblete, and C. Castillo (2010) 'Twitter under Crisis: Can We Trust What We RT?', in *Proceedings of the First Workshop on Social Media Analytics (SOMA '10)*, 71-79. Retrieved 13 Apr. 2014 from http://snap.stanford.edu/soma2010/papers/soma2010_11.pdf.

O'Brien, J. (2011, 22 Feb.) 'Removal of Export and Download / API Capabilities', *Archive of TwapperKeeper Blog*. Retrieved 1 Sep. 2014 from http://twapperkeeper.wordpress.com/2011/02/22/removal-of-export-and-download-api-capabilities/.

Papacharissi, Z., and M. de Fatima Oliveira (2012) 'Affective News and Networked Publics: The Rhythms of News Storytelling on #Egypt', *Journal of*

Communication, 62, 266-282. doi:10.1111/j.1460-2466.2012.01630.x

PR Newswire (2014, 29 Jan.) 'Facebook Reports Fourth Quarter and Full Year 2013 Results'. Retrieved 1 Sep. 2014 from http://www.prnewswire.com/news-releases/facebook-reports-fourth-quarter-and-full-year-2013-results-242637731.html.

Puschmann, C., and J. Burgess (2014) 'The Politics of Twitter Data,' in K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann (eds.), *Twitter and Society* (New York: Peter Lang), 43-54.

Rogers, R. (2014). 'Foreword: Debanalising Twitter: The Transformation of an Object of Study', in K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann (eds.), *Twitter and Society* (New York: Peter Lang), ix-xxvi.

Twitter, Inc. (2014) 'About'. Retrieved 1 Sep. 2014 from https://about.twitter.com/company.

Twitter, Inc. (2013, 2 July) 'Rules of the Road'. Retrieved 1 Sep. 2014 from https://dev.twitter.com/terms/api-terms.

Twitter, Inc. (2012, 25 June) 'Terms of Service'. Retrieved 1 Sep. 2014 from https://twitter.com/tos.