Easy Data, Hard Data: The politics and pragmatics of Twitter research after the computational turn

Jean Burgess and Axel Bruns

The implications of the computational turn

In the agenda-setting introductory chapter of his edited collection *Understanding Digital Humanities*, David Berry (2012) argues that the contemporary conjuncture in the history of the digital humanities constitutes a "computational turn." In Berry's formulation, this concept captures an historical moment that has both pragmatic and political dimensions. Berry defines it in terms of the use of computational technologies by disciplines in the humanities and social sciences, not only to "collect and analyse data with an unprecedented breadth and depth and scale," as Lazer et al. (2009) put it in their definitive statement on computational social science, but also "to shift the critical ground of [these disciplines'] concepts and theories" (Berry 2013, 11).

As the duality of this definition already shows, Berry's elaboration on the computational turn is much more than a procedural description of matters of fact affecting the self-identified digital humanities field. Rather, it is actually a provocative call to action directed at the humanities more broadly. Indeed, Berry draws on Presner's (2010) dramatic characterisation of this situation according to which we were "at the beginning of a shift in standards governing permissible problems, concepts, and explanations, and also in the midst of a transformation of the institutional and conceptual conditions of possibility for the generation, transmission, accessibility, and preservation of knowledge" (Presner 2010, 10). While cautioning humanities scholars not to leave the design of research tools and environments to private industry, and noting that it is vital to examine the cultural assumptions and constraints encoded into

both consumer- and programming-level digital technologies from Facebook to XML (10), Presner was clearly optimistic about Wikipedia as a model for digital transformations of the very structures of knowledge (11). At this point—four years on and counting—we note that in the context of an increasingly "platformed" (van Dijck 2013, 5–7) digital media environment that is "easier to use but more difficult to tinker with" (6), Presner's "transformation of the institutional and conceptual conditions of possibility" may entail as much closing down as opening up of scientific opportunities.

Taking these further complications into account and translating Berry's schema for understanding the implications of the computational turn in the field of sociocultural digital media research, the present chapter works from the premise that the computational turn has two main dimensions, which, taken separately and together, have significant implications for the competing futures of social media research in the humanities and social sciences.

The first dimension concerns the *practical* opportunities presented to humanities and social science researchers by computational approaches and methods. Here we are especially interested in the opportunities presented by the massive uptake and use of digital platforms for communication, the public availability of social media data that is generated by these activities, and new methods being developed to tackle the challenges of scale that result. In other words, this dimension describes the ways that computation can help to accomplish the pragmatic goals of empirical sociocultural research. It is no accident that in following this thread of analysis in this chapter, we focus almost exclusively on Twitter research. Indeed the comparative openness of Twitter as a platform (from both a user and a developer perspective) has both produced an unprecedented wealth of social media data and stimulated rapid and significant innovation in computational tools to gather, analyse, and visualise this data in the

digital humanities and social sciences (see for example Weller et al. 2014). This openness has also, accordingly, provoked transformation and challenge in the institutional structures and conceptual frameworks of sociocultural digital media research (Rieder & Röhle 2012, Burgess & Bruns 2012).

The second dimension of the computational turn concerns the application of critical humanities and social science theories and methods to the material *politics* of computational culture. In the context of data-driven digital media research, this dimension entails an additional layer of reflexivity, because it requires critical approaches to the computational aspects of social media platforms themselves. For the purpose of this chapter, the second dimension requires that we attend to the means by which platforms shape and control access to the data we use to diagnose the patterns of personal and public communication that they mediate. Here we can point to the growing body of work we might variously describe as "software studies" (Fuller 2008; Helmond 2013; Bucher 2013), "platform studies" (Bogost & Montfort 2009), or the "politics of platforms" (Gillespie 2013a) as a guide, with specific work on sociotechnical elements of Twitter (van Dijck 2013) and Facebook (Gerlitz & Helmond 2013; Bucher 2012) as exemplars, as well as earlier work reflecting on the politics of empirical research on social media platforms (Langlois & Elmer 2012). In a parallel set of considerations, José van Dijck argues that a truly materialist approach to the politics of social media platforms—the "techno-cultural constructs" (2013, 29) that constitute so much of the digital media environment—increasingly requires the combination of actor-network theory and political economy approaches (2013, 26–28). In what follows, we draw on both, albeit implicitly.

For digital media researchers working directly with social media data (whether big or small), these political and pragmatic dimensions of the computational turn are not

easily separated. Given that we will always be at least affected by, if not actively engaged in, *both* of these dimensions in the simple act of carrying out our work, it becomes clear that the computational turn is already starting to have some fairly profound transformative effects on the field of digital media research. It not only transforms how we go about doing media and communication studies—for example, introducing into the field automated methods of data collection and computational techniques of analysis and visualisation, redrawing and blurring disciplinary boundaries between the humanities, social sciences, and computer science; it also reconfigures the field itself, enlarging the scope of our potential objects of study—for example, requiring us to apply our analytical techniques to the study of the governing functions of APIs and the mechanics of hashtags, and not only the textual units of meaning they carry.

In the following section of the chapter, we draw out the relevant themes from a range of critical scholarship from the small body of digital media and software studies work that has focused on the politics of Twitter data and the sociotechnical means by which access is regulated. We highlight in particular the contested relationships between social media research (in both academic and non-academic contexts) and the data wholesale, retail, and analytics industries that feed on them. In the second major section of the chapter we discuss in detail the pragmatic edge of these politics in terms of what kinds of scientific research is and is not possible in the current political economy of Twitter data access. Finally, at the end of the chapter we return to the much broader implications of these issues for the politics of knowledge, demonstrating how the apparently microscopic level of how the Twitter API mediates access to Twitter data actually inscribes and influences the macro level of the global political economy of science itself, through re-inscribing institutional and traditional disciplinary privilege

We conclude with some speculations about future developments in data rights and data philanthropy that may at least mitigate some of these negative impacts.

Platform politics and regimes of access

Each of the currently-dominant, proprietary social media platforms is profoundly political. They are sociotechnical institutions that operate at multiple levels—technically, rhetorically, and culturally (Gillespie 2013)—to coordinate social and cultural interactions and expressions. As van Dijck reminds us, platforms are active mediators rather than neutral intermediaries, shaping "the performance of social acts instead of merely facilitating them" (2013, 29). At the same time, platforms are themselves shaped by use, and by the ways that they are represented in media, academic, and policy discourse. For researchers situated within sociocultural traditions but working with "big social data" (Manovich 2010) from social media, the politics of platforms are therefore ever-present as we go about what might otherwise appear to be purely instrumental exploitations of their affordances.

While a number of scholars are turning their critical attention to platforms understood in this way, the politics of the situation are often framed quite crudely—as a set of unilateral power relations between the platform, understood as a singular entity, and users, understood as a fairly homogeneous mass. But given the society-wide uptake of social media for a range of communicative purposes, it is important to differentiate these actors and the relations among them more carefully. We do so in this chapter by focusing on the politics of a particular set of relations—those that make up the field of computationally-assisted academic research relying on Twitter data—around a specific issue: the scientific affordances of Twitter data and the regulatory regimes of access that govern such uses. Thus we bring into the picture the relations among the platform

provider Twitter, Inc., the Twitter APIs, the third-party tools for accessing, analysing, and sharing data, and the ancillary economies of data mining and data retail that have sprung up around these affordances, as well as a range of commercial and not-for-profit research institutions and individual scientific users of the Twitter API for research purposes. In doing so, we make explicit the material politics of social media data access and its sociotechnical regulation as these relate to the wider critique of the politics of platforms (Gillespie 2013a; 2013b).

Part of the gradual but inexorable "making over" of Twitter from a social networking utility to a major media company (Burgess 2014) has been realised materially in changes to the technical and regulatory mechanisms that enable and constrain access to Twitter data, both for software development and research purposes. As Puschmann and Burgess (2014) have already discussed, social media data itself has become quickly commodified. Intermediaries who literally make it their business to package, sell, and analyse the collective interactions and expressions of millions of global social media users are emerging at a rapid rate. At the same time, the Twitter APIs have become less and less friendly to free-range data uses, including those of the third-party developer community that is arguably responsible for many of the innovations that give Twitter its unique culture, from the @reply and the hashtag to the search function and trending topics (Halavais 2014; Bruns 2012). But Twitter today is in the media business. It has been deliberately driven further and further away from being the neutral, transparent, and endlessly configurable "utility" that, back in 2009, co-founder Jack Dorsey famously declared he wanted it to be (McCarthy 2009) and toward being a managed advertising and interoperable media platform with a tightly controlled brand and user experience (Bilton 2012).

In this shift, the future monetisation of data analytics—including scientific data analytics—drawing on social media data is clearly seen as an important commercial growth area. Jose van Dijck has framed this shift as one from a situation where the possible meanings and uses of Twitter were many and varied—which, borrowing from Bijker et al., she characterizes as a phase of great "interpretative flexibility" (2013, 68–69)—to a reduction in this flexibility, a pattern of enclosure marked in 2011 by new restrictions on and governance of API access (84–85). As van Dijck correctly deduces, the regimes of access to Twitter data have become more clearly articulated and more tightly managed in order to enable the monetisation of data mining, which requires Twitter to have some controls over the entire pipeline of data, from the moment a tweet is uttered onward. Twitter's quasi-monopolistic licensing arrangements with, and then corporate acquisition of, data retailer Gnip in April 2014 (Messerschmidt 2014) are clear and relevant examples of Twitter's frequently remarked-upon corporate transition from a relatively free-range innovation platform to a much more managed, media-centric one, particularly since 2011 (Bilton 2012; 2013).

These significant changes in the ideologies and business models of social media platforms have been accompanied over the past five years by the increased "algorithmization" (Helmond 2013) of the social web, including the very techniques and architectures that social media platforms run on—from search to content discovery to target behavioural advertising and beyond. Famously, the delegation of cultural and knowledge work like content curation, sorting, classification, and evaluation for fuzzy concepts like "relevance" is increasingly being delegated to algorithms (Gillespie 2013b)—from Netflix's user preference predictions (Hallinan & Striphas 2014) through to Google's search results (Hillis et al. 2013) to the Facebook newsfeed (Bucher 2012). Our culture and society are increasingly mediated not only by knowable, single

algorithms (as the Google Search algorithm may once have been), but now by mobile and dynamic assemblages of algorithms, creating an almost incomprehensible situation (even for the engineers who build particular parts of the algorithms, or for the companies who hope to profit from them). As Gillespie (2014) argues, the algorithmic turn is stretching the "black box" metaphor so loved by STS scholars well beyond its limit. Whether truly knowable or reverse-engineerable by the platform providers, it is only the platform providers that have what Gillespie (2014) evocatively calls "backstage access" to "the public algorithms that matter so much to the public circulation of knowledge" (185). We suggest the same is true of the protocols and interfaces (such as APIs) that enable and govern public access to the data that constitute public communication after the computational turn.

While the cultures and meanings of Facebook and even Twitter are substantially and increasingly shaped by the algorithms used to search, curate, and suggest content to users, it is above all the Twitter APIs that play the most significant role in the mediating function of Twitter as a platform for the purposes of third party development, user innovation, and scientific research. But in comparison to, say, Facebook's newsfeed algorithms, there is far less critique of the politics of APIs. One notable exception here is Taina Bucher's (2013) study of the changing politics of Twitter's APIs, which includes valuable interview material garnered from conversations with members of the third-party developer community, demonstrating how APIs (including the Search and Streaming APIs that are essential to Twitter research) work as "quasi-objects" mediating the relations between actors and uses, and evoking "intense feeling" within the developer community.

But changes to the Twitter API have affected the research community as well. Bruns and Burgess (forthcoming) have traced the ways that Twitter's changing business

model, as instantiated through the API and associated rule changes, has impacted Twitter research over time. Gradual changes to what forms and volumes of data are available through the API, and at what retrieval speeds, have directly impacted academic research on the uses of the platform. For example, while during the early years of the platform researchers were able to request and gain "whitelisted," premium access to the API at the discretion of Twitter support staff, enabling them to retrieve a larger amount of data at greater speeds, such access (or the equivalent thereof) is now available only to the paying subscribers of third-party Twitter data resellers (such as Gnip or DataSift) and at significant cost, creating a "missing middle" in terms of data accessibility and putting small-scale data access facilitators out of business, at least where Twitter data is concerned (Irving 2014). Instead, Twitter's business model has relied progressively more on providing licensed data access to large players, in keeping with the rise of the "social data" market (Puschmann & Burgess 2012), further complicating not only the practical accessibility of the platform to researchers, but introducing new dimensions of the politics of such research.

This research becomes further entangled with the politics of platforms as social anxieties over corporate data mining and government surveillance mount. And matters are further complicated when corporate research crosses over into the public domain, raising serious questions about scientific research ethics after the computational turn. The most striking example here is the misnamed "emotional contagion" experiment run on the Facebook newsfeed without obtaining informed consent and without debriefing those users whose Facebook feeds were part of the experiment (Kramer et al. 2014; see also the follow-up on ethics from Kahn et al. 2014). As this controversy demonstrated, increasingly the tools, techniques, and methodologies of corporate data mining are becoming less distinct from academic ones; the same cannot necessarily be said of

epistemological foundations, research motivations, and ethical considerations. The conversation on social media research ethics within the academic digital media research community is only just beginning (Zimmer & Proferes 2014) and is notably muted in contexts where large-scale Twitter data plays an important role and research teams might be operating in grey areas with respect to compliance with the Twitter Terms of Service, for example. Here we should note that privacy concerns are substantially integrated into the regulatory controls that Twitter, Inc. is deploying within its data retail business through Gnip. Simply handing over large amounts of money to Gnip is insufficient to gain access; rather, projects have to be comply with Twitter's protocols for data use, instantiated in the licensing agreement between them as platform provider and Gnip as data reseller and additionally assessed on a case-by-case basis (Irving 2014), but not publicly available on the Gnip or Twitter websites.

We now turn to a more detailed narrative account of how scientific researchers—particularly those in the humanities and social sciences—have been using Twitter data, the way these regimes of access have shaped the kinds of methods and approaches that are possible (or even the kinds of questions that can be asked), and how changes to these regimes of access are introducing new challenges and limitations to such research. In so doing, we hope that a reflexive account from the perspective of researchers themselves can at least open up some of the "backstage" questions about the politics of Twitter data, if not their answers, as we reflect on the material aspects of the research challenges presented by attempts to negotiate data access and use via the Twitter APIs, Terms of Service, and disciplinary norms and expectations.

The changing scientific affordances of Twitter data

As we have already outlined, Twitter's early history is characterised by—and indeed in many ways driven by—the existence of a strong and productive network of third-party developers and researchers providing a range of enhancements to the standard Twitter experience. Such enhancements included a variety of stand-alone Twitter clients that were more user-friendly than, or provided different functionality from, the standard web-based Twitter user interface or Twitter's own smart phone apps—and even at this early stage often already offered some basic Twitter analytics, designed mainly to enable users to trace the trajectory of their own Twitter presence. On its own, the emergence of this ecosystem of Twitter enhancements can also be seen simply as the logical next stage of a history of platform co-design and co-evolution processes that for Twitter has always involved the company Twitter, Inc. and its lead users; even what are now regarded as some of the most basic interventions in Twitter's array of communication technologies—the @reply and the hashtag—were after all introduced by Twitter users and only subsequently incorporated into Twitter's core system (Halavais 2014; Bruns 2011).

As a fledgling social media service, emerging in early 2006 alongside the already well-established Facebook, and not receiving any substantial user and media recognition until early 2009, Twitter initially had a strong vested interest in facilitating widespread experimentation with its communicative features, and in working with third-party developers (themselves largely early adopters and enthusiasts) as they explored the technical and commercial viability of particular add-ons, thus identifying potential new applications for Twitter as a social media service. After all, any new uses for Twitter as a platform and, in particular, any "killer app" features that might position

Twitter as an attractive alternative to Facebook, would also contribute considerably to Twitter's own viability as a platform.

Throughout these early days of Twitter's history, Twitter, Inc. therefore provided significant support and encouragement to the ecosystem of developers that gravitated to its platform. Central to such support was the provision of a relatively open and powerful API which provided access to user profile and tweeting activity data. Notably, and owing also to the comparatively flat and open structure of the Twitter network itself, the Twitter API was considerably more powerful and less restricted than its Facebook counterpart. And while some restrictions did apply to the volume of data that ordinary API clients could access, Twitter, Inc. also operated a comparatively generous regime of API client "whitelisting," removing most of these restrictions for API users whose activities were considered *ad hoc* and of some value or interest to Twitter and its users.

While third-party Twitter developers made up the majority of the recipients of such whitelisted access, a number of scholarly and commercial Twitter researchers also benefitted from such generosity, and even the non-whitelisted API proved highly attractive to Twitter researchers. This is reflected in the substantially greater number of scholarly research papers published to date that draw on the Twitter rather than Facebook API: in late 2014, a simple Google Scholar search for "Twitter" and "API" returns more than 560,000 results, compared to only 158,000 for "Facebook" and "API"). Indeed, much as the powerful Twitter API provided a general boost to the development of Twitter clients and other enhancements, it also led to the development of a number of widely used specialist Twitter research tools and applications.

The early tools for conducting Twitter research by gathering data from the API largely inherited the same spirit of enthusiasm and exploration as their generic

counterparts; alongside the first Twitter researcher get-togethers under the auspices of leading academic conferences such as the annual Association of Internet Researchers conference, or in stand-alone events such as the Düsseldorf Workshop on Interdisciplinary Approaches to Twitter Analysis (DIATA), they contributed significantly to the emergence of a global network of Twitter researchers who shared their methods, approaches, and findings, and often also their data sets, with each other. Indeed, platforms such as *Twapperkeeper.com*—one of the most popular early tools for gathering Twitter data by tracking and archiving tweets that contained specific user-definable keywords or hashtags—explicitly encouraged such data sharing by making their archives publicly available not just to the originating researcher, but to all other users of the platform.

However, in their encouragement of such collegial and accountable research practices, which enabled researchers to test and verify each other's findings by working with these shared data sets, these tools also increasingly ran afoul of Twitter, Inc.'s stated Terms of Service for the API. In April of 2011, the company forced *Twapperkeeper.com* to shut down its public service (O'Brien 2011). Twapperkeeper functionality was instead made available only as part of the open-source package *yourTwapperkeeper*, which requires researchers to install the tool on their own web servers and, in line with Twitter's Terms of Service, discourages the public sharing of the data sets it produces. This intervention by Twitter, Inc. thus had a considerable impact on the Twitter research community, both by limiting research activities to those researchers who had the capacity to operate their own data-gathering servers and by undermining researchers' ability to (legally) share and verify each other's data sets. Virtually all Twitter research tools and facilities that have emerged in the meantime (see Gaffney & Puschmann 2014, for a useful overview) similarly require researchers to

operate their own servers, and provide data-gathering functionality that is broadly comparable to that of *yourTwapperkeeper*.

Subsequent to its shutdown of *Twapperkeeper.com*, Twitter, Inc. also began to circumscribe the functionality of the various elements of its Application Programming Interface considerably more tightly. Most centrally, it introduced a range of limits seriously restricting the number of calls to the user and search APIs (which provide information on user profiles and past tweets, respectively) that a Twitter client could make in each 15-minute window. Additionally, the search API also provides no more than the last 3200 tweets for any given user or search term, while the streaming API (which delivers a continuous feed of tweets matching a given search term) only delivers up to one per cent of the total current throughput of the full Twitter fire hose—that is, of the full feed of all current tweets. Further, unrestricted access to the data, which these APIs are designed to provide (that is, to older historical tweets or higher-volume current content) can now no longer be obtained by applying for whitelisted access. Rather, the only generally available method for gaining such access is to buy data at a substantial cost from one of a handful of authorised commercial Twitter data resellers, thus placing larger-scale access to Twitter data out of reach of most publicly-funded research projects and institutions.

By contrast, what remains easily accessible through the standard, open APIs to Twitter researchers without the funds to buy data access are the Twitter feeds that can be generated by tracking a set of user-defined keywords or hashtags, provided that such feeds return a combined total volume of tweets that remains below the one per cent limit that applies to the streaming API. The net effect of such access-shaping policies has been to push the emerging field of Twitter research to by and large focus on a relatively narrow range of research questions and comparatively isolated case studies, resulting

in a dominant form of Twitter research that may be summarised without over-simplification as "hashtag studies," since it largely gathers and analyses Twitter data sets defined by the presence of one or more set hashtags. Such hashtag-centric research is undoubtedly valuable in its own right—it has variously shed light on the uses of Twitter in crisis communication (Palen et al. 2010; Mendoza et al. 2010), political controversies (Maireder & Schlögl 2014; Hermida 2014), and second-screen viewing (Highfield et al. 2013), for example—but inevitably fails to place its findings in a broader, Twitter-wide context, since the very design of Twitter's API restrictions makes it virtually impossible for researchers to gather such contextual information. Indeed, even the one per cent limit of the streaming API, which may affect data gathering for particularly active hashtags and keywords, is rarely recognised and problematised in the methodological discussion of hashtag studies papers.

Such a contextualisation of hashtag studies in the wider context of Twitter activity is necessary and crucial. While it is useful, for example, to examine the volume and dynamics of user activities within a hashtag data set on a major crisis, such data sets miss out by design on any of the further ancillary public communication activities that happen around them, sparked by or leading to new hashtagged tweets. For example, users responding to a hashtagged tweet may themselves not include the hashtag in their responses, since they are now directly @replying to another user. Conversely, users may afford greater visibility to a tweet they encountered through their own network by retweeting it with an added topical hashtag. But neither of these subsequent or preceding activities will be visible in a hashtag data set. Similarly, while the volume of tweets captured in the data set may appear substantial, its full significance can be assessed only against the benchmark of the total volume of Twitter activity at the same time. Finally, the mere collection of hashtagged tweets itself provides very little

indication of its likely impact on public communication on Twitter unless researchers can also draw on detailed information about the participating accounts and their positioning in the network of followers and followees: a hashtagged conversation may have taken place entirely between already tightly connected accounts, for example, or have brought together complete strangers; or it may have involved many virtually friendless accounts or a number of Twitter's most influential users. The contextual data underlying such further analysis is considerably harder to gather than the hashtagged data sets themselves; for a given hashtag data set, it would require the researcher to identify all participating accounts, and then to retrieve the public profile, recent tweeting history, and follower network information available for each of these profiles. Due to the significant limitations that apply to the relevant API access points for large hashtag data sets, this may be a slow and laborious undertaking.

Further, the continuing centrality of the hashtag data set as the first stage in any such research process also maintains an implicit assumption that hashtags do indeed continue to act as a core coordinating principle for Twitter interactions. But while this may be true for specific acute events (Burgess & Crawford 2011) and other forms of breaking news that lead to the rapid formation of an *ad hoc* public (Bruns & Burgess 2011) around emerging hashtags, that assumption is unlikely to be sustainable for most other forms of Twitter interaction. By contrast, the most common—and arguably least researched—form of public communication through Twitter is everyday interaction between users who simply encounter each other's posts because of unidirectional or reciprocal follower/followee relationships. Outside of crisis events, the routine monitoring of the Twitter feed, comprising all recent tweets made by the accounts followed by a given user, must clearly be considered the standard mechanism through which information is disseminated across the network.

To research the dynamics of such information dissemination and user interaction, then, would require a very different combination of data sets than is provided to users of the standard, open Twitter API. At minimum, researchers would need to define a population of Twitter accounts to be monitored; they would have to gather the follower/followee network information for all of these accounts; and they would have to monitor their public communication activities on an ongoing basis for the duration of the project. Most crucially, they would have to do so for a relatively large population of Twitter accounts (for example, all of the accounts in a specific region) in order to generate any meaningful insights that are not inherently skewed by the particular population of accounts tracked. In principle, any such data are also available from the open Twitter API; in practice, however, the API restrictions would make gathering them a slow and laborious process. Further, due to the continuing popularity of hashtag studies in the scholarly community, most of the data-gathering, -processing, and -analysis tools required for such work would have to be developed from scratch: the emphasis on hashtag studies, caused by Twitter, Inc.'s shaping of API access to privilege specific forms of data gathering over others, is thus self-reinforcing.

In essence, then, unless they can raise the significant funds required to buy data from commercial reselling services, Twitter researchers are today forced to limit their activities to working with the "easy" data that are readily available through the standard Twitter API services, or to engage with a clandestine network through which specific Twitter data sets, as well as methods and tools for circumventing the API access limits, are now being exchanged. Meanwhile, the "hard" data that require higher-level Twitter access and more powerful research facilities have become the sole domain of commercial research institutions and a handful of well-resourced research labs,

which—often in collaboration with commercial partners—have managed to raise the funds required to buy access to the Twitter fire hose or similarly large data sets.

This has led to a certain stagnation in the development of Twitter research. Since the majority of Twitter researchers now emerging are still coming to terms with the research methods and approaches that are available to them, the chilling effects of a reduction of Twitter studies to mere hashtag studies are not yet widely appreciated and articulated. But the chilling effects of Twitter, Inc.'s restrictive policies do have a significant detrimental effect on both the quality and diversity of scholarly research investigating the uses of Twitter as a major medium for public communication today—and this ultimately not only affects the scholarly community and the public debates it contributes to and engages with, but also Twitter itself, as both a platform and a company. If very little intellectually hard research on Twitter is able to draw on large data sets, and most of the technically hard, "big data" Twitter research is conducted by commercial and market researchers, our understanding of Twitter as a media platform and of Twitter-based communication as a sociocultural phenomenon must necessarily suffer.

Conclusions

In this chapter, we have laid out the dual dimensions of the "computational turn" and how they become entangled with the practice of digital media research in the humanities and social sciences. We have focused particularly on some uses of critical software and platform studies for dealing with these entanglements at the very specific and concrete level of understanding how the Twitter API mediates access to Twitter data, and how changes to these "regimes of access" relate to the changing business realities and aspirations of Twitter, Inc., as well as to the "datafication" and

"algorithmization" of the digital media business more generally. We focused particularly on the affordances of Twitter data for academic research and how these have been shaped and constrained by Twitter's regimes of access to data. Here we would like to emphasise that it is not only wholesale access per se, but also the differential availability of data—ranging from the various parts of the tweets and their metadata in hashtag and keyword data sets through to more comprehensive Twitter feeds, public profile information, and follower-followee networks—that shape the research questions that can be asked. Such constraining factors, resulting from the business decisions of Twitter, Inc., shift the emphases of the entire emerging field of "Twitter studies," which today tends to focus too heavily on @reply networks and hashtag publics because current API limitations privilege certain forms of access and use over others.

We therefore distinguish between "easy data" on the one hand—modestly sized sets of tweets and certain associated, pre-determined metadata matching a keyword search over a short, recent period of time—and "hard data" on the other—more comprehensive, longitudinal data sets and/or any of the "missing" metadata. We have shown how access to the hard data bestows very considerable scientific advantages on those who have it, and suggested that it is problematic if there is a correspondence between scientific privileges of this kind and the prestige, wealth, and industry-connectedness already enjoyed by elite universities and research institutes—particularly in those cases where such research is conducted behind closed doors, to the benefit of industry but not to the collective benefit of the platform's differentiated user community.

On this basis we argue that there is now a growing divide between the majority of researchers who are forced to work with easy data—pursuing the low-hanging fruit in Twitter research because a lack of funding for data access and research tools

development prevents them from doing anything else—and a minority of researchers and institutes who either have the funds to pay Gnip or DataSift, and whose research complies with the opaque corporate restrictions Twitter places even on its data-selling partners, or who have the technical skills to partially circumvent API restrictions and access the more difficult "hard data" about public communication on Twitter. Due to underlying differences in funding patterns, this divide is also one between different disciplinary perspectives: prominent commercial as well as university-based market research and computer science institutes dominate in the latter group (including the research labs associated with some of the world's biggest technology companies), while their poorer cousins from the humanities and social sciences, as well as from underfunded universities or those in developing countries without large research budgets, are more likely to be found in the former.

Whether intended by Twitter, Inc. or a collateral outcome of internal priorities, this growing imbalance significantly skews the trajectory of Twitter research, and thereby is likely also to affect the public perception of Twitter as a tool for public communication. Researchers who wish to dig deeply into the political economy of social media platforms—including those who wish to take up van Dijck's call to integrate actor-network theory and political economy approaches when doing so—would do well to consider the regimes of access, politics of use, and economies of exchange that increasingly constitute the market in "big social data" research. Researchers more pragmatically concerned with access to Twitter data and the full range of (still circumscribed) scientific possibilities it affords—even if frustrated in their attempts to gain access—may nevertheless contribute to the global politics of knowledge, through public reflections and critiques of the regimes of access that emerge in relation to scientific uses of social media data.

Finally, while beyond the scope of our discussion in this chapter, it is possible that the politics of data access for Twitter and other social media platforms may have important legal and policy ramifications as well, especially in light of the growing "data rights" and information rights movements to which the growing debate around "data philanthropy" is contributing (see for example Stempeck 2014). Twitter's donation of the full historical Twitter fire hose to the US Library of Congress in 2010, as well as 2014's inaugural Twitter data grants program (Kirkorian 2014) are framed as examples of Twitter's largesse in data philanthropy; even third-party data reseller DataSift has announced humanitarian data science "partnerships" involving gifted data access (DataSift 2014).

Such grants are welcome, in principle, if they are truly philanthropic rather than merely an exercise in corporate public relations. The real impact of such initiatives to date is entirely negligible: the Library of Congress Twitter archive has yet to be made available to researchers in any meaningful and transparent fashion, and Twitter data grants were awarded to a paltry six projects from a total field of more than 1,300 applicants (Kirkorian 2014). Developments such as these should give us pause: why is it that commercial data retail companies are selectively making data "grants" to one-off projects (by proxy taking over the evaluative role of national science agencies and global not-for-profit organisations in selecting worthy initiatives to support), rather than developing widely accessible, affordable models of access to the public's data for public good? Given the strong demand for sustainable scientific access to social media data sets beyond what can be retrieved through their open APIs, it would be more responsible and appropriate for Twitter, Inc. and other providers to develop meaningful and affordable data access frameworks and pricing structures that address the growing

missing middle in scholarly social media research, rather than operating what in essence amounts to a random data lottery.

References

Berry, David. 2012. "Introduction: Understanding the Digital Humanities." In David Berry (Ed.), *Understanding Digital Humanities*. London: Palgrave Macmillan.

Bilton, Nick. 2013. *Hatching Twitter: A True Story of Money, Power, Friendship, and Betrayal*. New York: Portfolio.

Bilton, Nick. 2012. "Is Twitter a technology company or a media company?" *Bits Blog (New York Times)* http://bits.blogs.nytimes.com/2012/07/25/is-twitter-a-media-or-technology-company/?_php=true&_type=blogs&_r=0

Bruns, Axel. 2012. "*Ad Hoc* Innovation by Users of Social Networks: The Case of *Twitter*." ZSI Discussion Paper 16. https://www.zsi.at/object/publication/2186.

Bruns, Axel, and Jean Burgess. 2014. (forthcoming). "Methodological Innovation in Precarious Spaces: The Case of Twitter." In Helen Snee & Yvette Morey (Eds.), *Digital Methods for Social Science*. London: Palgrave Macmillan.

Bucher, Taina. 2013. "Objects of Intense Feeling: The Case of the Twitter APIs." *Computational Culture* 3. http://computationalculture.net/article/objects-of-intense-feeling-the-case-of-the-twitter-api

Bucher, Taina. 2012. "Want to Be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook." *New Media & Society* 14(7), 1164-1180.

Burgess, Jean. 2014. "From 'Broadcast yourself' to 'Follow your interests': making over social media. *International journal of cultural studies* [online first]. http://ics.sagepub.com/content/early/2014/01/13/1367877913513684.abstract.

Burgess, Jean, and Axel Bruns. 2012. “Twitter Archives and the Challenges of ‘Big Social Data’ for Media and Communication Research.” *M/C Journal* 15(5). http://journal.media.culture.org.au/index.php/mcjournal/article/viewArticle/561

Bruns, Axel, and Jean Burgess. 2011. “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. http://eprints.qut.edu.au/46515/.

Burgess, Jean, and Kate Crawford. (2011). “A Theory of Acute Events in Social Media.” Paper presented at the Association of Internet Researchers conference, Seattle, 11-13 Oct. 2011.

DataSift. (2014, 2 July). “UN Global Pulse & DataSift Announce Data Philanthropy Partnership.” http://datasift.com/press-releases/UN-Global%20Pulse%20Partnership/.

Fuller, Matthew. 2008. *Software Studies: A Lexicon*. Cambridge, MA: The MIT Press.

Gaffney, Devin, and Cornelius Puschmann. 2014. “Data Collection on Twitter.” In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann, eds., *Twitter & Society*. New York: Peter Lang, 55-68.

Gerlitz, Carolyn, and Anne Helmond. (2013). “The Like Economy: Social Buttons and the Data-Intensive Web.” *New Media & Society* 15(8): 1348-1365.

Gerlitz, Carolyn, and Bernhard Rieder. 2013. “Mining One Percent of Twitter: Collections, Baselines, Sampling.” *M/C Journal* 16(2). http://www.journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/620

Gillespie, Tarleton. 2013a. “The Politics of ‘Platforms’.” *A Companion to New Media Dynamics*, Eds John Hartley, Jean Burgess and Axel Bruns. London: Wiley-Blackwell, pp. 407-16.

Gillespie, Tarleton. 2013b. “The Relevance of Algorithms.” *Media Technologies*, Eds Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot. Cambridge: MIT Press.

Halavais, Alex. 2014. “Structure of Twitter: Social and Technical.” In Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann, Eds., *Twitter & Society*. New York: Peter Lang, 29-42.

Helmond, Anne. 2013. “The Algorithmization of the Hyperlink.” *Computational Culture* 3. http://www.annehelmond.nl/wordpress/wp-content/uploads/2013/11/Helmond_2013_CC_AlgorithmizationOfTheHyperlink.pdf

Hermida, Alfred. 2014. “Contested Media Spaces: #idlenomore as an Emergent Middle Ground.” Paper presented at *Social Media and the Transformation of Public Space*, University of Amsterdam, 19 June 2014.

Highfield, Tim, Stephen Harrington, and Axel Bruns. 2013. “Twitter as a Technology for Audiencing and Fandom: The #Eurovision Phenomenon.” *Information, Communication & Society* 16(3): 315–39.

Hillis, Ken, Petit, Michael, and Kylie Jarrett. 2013. *Google and the Culture of Search*. New York: Routledge.

Irving, Francis. (2014). “The Story of Getting Twitter Data and its ‘Missing Middle’.” *ScraperWiki Blog*. https://blog.scraperwiki.com/2014/08/the-story-of-getting-twitter-data-and-its-missing-middle/

Kahn, Jeffrey P., Effy Vayena, and Anna C. Mastroianni. 2014. "Opinion: Learning as We Go: Lessons from the Publication of Facebook's Social-Computing Research." *Proceedings of the National Academy of Sciences* 111(38): 13677-13679.

Kirkorian, Raffi. 2014, 17 Apr. "Twitter #DataGrants Selections." *Twitter Engineering Blog*. https://blog.twitter.com/2014/twitter-datagrants-selections.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences*, 111(24): 8788–8790.

Langlois, Ganaele, and Greg Elmer. 2013. "The Research Politics of Social Media Platforms. *Culture Machine*, 14. http://www.culturemachine.net/index.php/cm/article/viewDownloadInterstitial/505/531

Lazer, David, et al. 2009. "Life in the Network: The Coming Age of Computational Social Science." *Science* 323(5915): 721-23.

Maireder, Axel, and Stephan Schlögl. 2014. "24 Hours of an #outcry: The Networked Publics of a Socio-Political Debate." *European Journal of Communication* (online first 2 Sep. 2014). doi:10.1177/0267323114545710.

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

McCarthy, Caroline. 2009. "Twitter Co-Founder: We'll have Made it When you Shut Up About Us." *CNet.* http://www.cnet.com/au/news/twitter-co-founder-well-have-made-it-when-you-shut-up-about-us/

Mendoza, Marcelo, Barbara Poblete, and Carlos 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. http://snap.stanford.edu/soma2010/papers/soma2010_11.pdf.

O'Brien, Joe. 2011. "Removal of Export and Download / API Capabilities." *Archive of TwapperKeeper Blog*. http://twapperkeeper.wordpress.com/2011/02/22/removal-of-export-and-download-api-capabilities/.

Palen, Leisa, Kate Starbird, Sarah Vieweg, and Amanda 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.

Presner, Todd. 2010. "Digital Humanities 2.0: A report on knowledge." http://cnx.org/content/m34246/1.6/?format=pdf

Puschmann, Cornelius, and Jean Burgess. 2014. "The Politics of Twitter Data." In Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, & Cornelius Puschmann (Eds.) *Twitter and Society* (pp. 43-54). New York: Peter Lang.

Rieder, Bernhard, and Theo Röhle. 2012. "Digital Methods: Five Challenges." In David Berry (Ed.), *Understanding Digital Humanities* (pp. 67-84). London: Palgrave Macmillan.

Rogers, Richard. 2013. *Debanalizing Twitter: The Transformation of an Object of Study*. http://www.govcom.org/publications/full_list/rogers_debanalizingTwitter_websci13.pdf

Stempeck, Matt. 2014. "Sharing Data is a Form of Corporate Philanthropy." *Harvard Business Review*. http://blogs.hbr.org/2014/07/sharing-data-is-a-form-of-corporate-philanthropy/

Hallinan, Blake, and Ted Striphas. 2014. “Recommended for You: The Netflix Prize and the Production of Algorithmic Culture.” *New Media & Society* [online first]. http://nms.sagepub.com/content/early/2014/06/23/1461444814538646.abstract

van Dijck, José. 2013. *Culture of Connectivity: A Critical History of Social Media* (Kindle Edition). New York: Oxford University Press.

Weller, Katrin, Axel Bruns, Jean Burgess, Cornelius Puschmann, and Merja Mahrt (Eds.). 2014. *Twitter and Society*. New York: Peter Lang.

Zimmer, Michael, and Nicholas Proferes. 2014. “A Topology of Twitter Research: Disciplines, Methods, and Ethics.” In *ASLIB Proceedings* 66(3).