PRE-PRINT VERSION

Published: Phillip Brooker , Julie Barnett , Timothy Cribbin & Sanjay Sharma (2015) 'Have we even solved the first 'big data challenge?' Practical issues concerning data collection & visual representation for social media analytics', in H. Snee et al. (Eds) *Digital Methods for Social Science: An Interdisciplinary Guide to Research Innovation*. London: Palgrave MacMillan. ISBN: 9781137453655.

Have We Even Solved the First 'Big Data Challenge?': Practical Issues Concerning Data Collection and Visual Representation for Social Media Analytics

Abstract

The present chapter explores the technical and computational processes through which social media data is shaped into research findings. The authors make this argument by depicting the effects of two practical issues - API rate limiting in Twitter data collection and the use of spatial mapping algorithms in visualising those data - on resulting analyses. Such issues are not problematic to social media analytic research; rather, they can be used as *resources* for helping to characterise and understand the data at hand. Hence, the authors work to demonstrate the value in incorporating these reflexive analyses of technical and computational processes into our accounts; to advocate *thinking in assemblages* as a requirement for making analytic claims with 'big' social media data.

Introduction

Thanks to an influx of data collection and analytic software, harvesting and visualising 'big' social media data[1] is becoming increasingly feasible as a method for social science researchers. Yet whilst there is an emerging body of work utilising social media as a data resource, there are a number of computational issues affecting data collection. These issues may problematise any conclusions we draw from our research work, yet for the large part, they remain hidden from the researcher's view. We contribute towards the burgeoning literature which critically addresses various fundamental concerns with Big Data (see boyd and Crawford, 2012; Murthy, 2013; and Rogers, 2013). However, rather than focus on epistemological, political or theoretical issues – these areas are very ably accounted for by the authors listed above, and others – we engage with a different concern: how technical aspects of computational tools for capturing and handling social media data may impact our readings of it. This chapter outlines and explores two such technical issues as they occur for data taken from Twitter.

Throughout the chapter, we demonstrate a perspective consistent with Procter *et al*.'s (2013) view that social researchers wishing to make sense of 'big' social media data should have sufficient knowledge of the underlying concepts of the computational methods and tools they are using, so as to be able to decide when and where to appropriately apply them. Furthermore, we take heed of boyd and Crawford's suggestion that 'When researchers approach a data set, they need to understand – and publicly account for – not only the limits of the data set, but also the limits of which questions they can ask of a data set and what interpretations are appropriate' (2012: 669-70). To this end, we highlight how certain technical characteristics and constraints pertaining to the collection and processing of Twitter data can impact on research and how an understanding of these factors might lead to more robust accounts of such data.

Our aim is to demonstrate the mainstream relevance of a commonplace methodological procedure in the social sciences; namely the self-critical reflexive analyses of our methods in terms of their impact on our accounts of the subjects we study. Our goal here is to show the importance of understanding the effects that technical processes may have on our readings of data for all social scientists, not just for those with a background in computer science. Without this understanding it is impossible to make sense of the data at hand. Hence, we promote the idea of thinking of the investigative process as an 'assemblage' (Langlois, 2011; Sharma, 2013) that draws together various social and technical (and other) factors into a unified research process. Here, we refer to the ways in which the research process comes to feature not only conceptual theoretical knowledge and inductive empirical findings, but also how technical issues (such as API rate limiting and spatial mapping algorithms) contribute towards the production of knowledge in multifarious complex ways. How such an assemblage might operate will become clearer as we present our selected two technical issues, and in the discussion that follows.

Reviewing the Field

The State of Social Media Analytics

With the field of social media analytics still in relative infancy there are few methodological practices taken as standard. The tendency thus far has been to fit digital data to existing 'offline' ways of working. As O'Connor *et al*. note:

> Often the online researcher has little in the way of research precedent to use as a guide to practice online research and, as a result, online researchers frequently turn to established offline practices (O'Connor *et al*., 2008: 276).

Working from this position, several authors characterise social media data as a special kind of 'offline' social science data. For example, Hine argues that a key concern of social media analytics is to avoid a loss of quality in data; 'Face-to-face interaction here becomes the gold standard against which the performance of computer-mediated interaction is judged' (Hine,

2006: 4). This quality problem of social media data is a concern shared by many, with comments being levelled at 'the lack of uniformity in how users fill in forms, fields, boxes and other text entry spaces' (Rogers, 2013: 205); the representativeness and validity of the data more generally (Tufekci, 2014); and the fact that the production of data is not controlled and led by researchers but appears untamed 'in the wild' (Kitchin, 2013). Kitchin notes:

> The challenge of analysing Big Data is coping with abundance, exhaustivity and variety, timeliness and dynamism, messiness and uncertainty, high relationality, and the fact that much of what is generated has no specific question in mind or is a by-product of another activity (Kitchin, 2014: 2).

Given the uncertain relationship between digital and 'offline' methods, it becomes important to explore possible ways of rendering visible the characteristics of digital methods to see how and where they fit into existing methodological practices. Our proposed treatment of such data embraces the 'digitality' of researching in this area by advocating a greater working familiarity with computational tools and methods. Emphatically, this is not to say that the work of social media analytics can be reduced to the rote operating of software (see Keegan (2013)). Kitchin summarises this tension:

> For many...the digital humanities is fostering weak, surface analysis, rather than deep, penetrating insight. It is overly reductionist and crude in its techniques, sacrificing complexity, specificity, context, depth and critique for scale, breadth, automation, descriptive patterns and the impression that interpretation does not require deep contextual knowledge. (Kitchin, 2014: 8)

Yet we do not believe this is *necessarily* how social media analytics has to operate. On the contrary, we advocate a mode of reading data that allows computational methods to pick out areas of potential interest which then might be explored more intimately through closer readings. To do this requires an understanding of how social media data can be affected by technical processes as part of a wider assemblage. As Lewis *et al.* note, 'As scholars increasingly take up such datasets, they may be better served by interweaving computational and manual approaches throughout the overall process in order to maximise the precision of algorithms and the context-sensitive evaluations of human coders' (2013: 49). By outlining how this kind of research process works 'on the shop floor', we hope to foster a way of thinking about such technical issues which might facilitate the mainstream usage of digital research methods generally.

We take steps towards respecifying 'big' social media work in this way by concentrating on two issues. Firstly we demonstrate the effects of a data collection issue – the rate limiting of Twitter's Application Programming Interfaces (APIs), which is a built-in restriction on the

flow of Twitter data that may interfere with analyses in significant ways. Secondly, we remark upon the ways in which computational models (and the visualisations that represent them) might shape our analytic readings of data. Those already working in the field are well aware of these concerns, yet they do not routinely feature in published accounts of relevant work. Consequently, such issues may stand as a barrier to entry by steepening an already steep learning curve. Hence, we openly discuss two such issues, not necessarily as problematic to social media analytics but as presenting an opportunity to make better use of new and powerful data resources.

Addressing the First 'Big Data Challenge'

Before doing so it is useful to describe what we are referring to in the title as the First 'Big Data Challenge'. One much vaunted promise of 'Big Data' is that we all now have the means to access data from sources like Twitter, and to engage in analytic work on large data corpora through processing that data into easily-digestible visualisations. Moreover, this work does not *necessarily* require any special skill with computer science or programming – there is a wealth of freely-available third-party software tools to do the 'behind the scenes' technical work for us[2]. In this sense, the First 'Big Data Challenge' refers to a) having easy access to big data, and b) the availability of tools that facilitate its analysis. Our question in the title – whether or not we are yet in a position to close the book on this first challenge – demonstrates our intention to probe such matters further: *can* we tap vast data resources unfiltered? Is it really as simple as employing visual models to show us the results? Such concerns are worked out through the process of *doing* social media analytics and acquiring necessary relevant skills along the way. Our approach here is to more accountably explore this process of doing social media analytics to help figure out what might count as appropriate methods and methodologies.

But why is it important to render transparent what might be argued to be mundane computational issues? Doing social media analytics with Twitter data necessitates an interfacing with the mechanisms governing how users access Twitter data: the Twitter Application Programming Interfaces (APIs). These APIs allow users to request to access certain slices of Twitter data, according to various search strategies (i.e. by keyword, by bibliographic/demographic information, by random selection, and so on). Moreover, once investigators have personal copies of these data, they may then subject them to further algorithmic processes to make their 'Big Data' analysable, e.g. in the rendering of statistical information or in the production of visualisations and so on. In this way, computational processes come to feature as essential elements in the production and construction of our data and analyses. As Marres notes, this bringing together of different disciplinary ideas can be equally productive and constricting:

> [Digital] social research is *noticeably* marked by informational practices and devices not of its own making, from the analytic measures built into online

> platforms (e.g. the number of links, number of mentionings, follower counts), to the visual forms embedded in visualization modules (the tag cloud). Online social research is visibly a distributed accomplishment (Marres, 2012: 160).

This intertwining of technical issues and research methods is foregrounded for the social sciences by Fielding and Lee, who argue that 'Social science has demonstrated how technology both shapes and is shaped by social action. Research methods are no exception' (2008: 505). As such, since there are computational processes governing data collection and analysis, we may find ourselves better-armed to undertake research in the area if we understand some of the finer points about how these tools and processes work. *How exactly* do they restrict the data we can harvest? And *how exactly* do they shape the statistics and visualisations and other analytic outputs which we use to understand that data?

To this end, we now turn to a more pointed examination of two such issues – the possible effects of API rate limiting on data collected through Twitter, and the possible effects of spatial mapping algorithms on data visualisations – as they occur through the usage of a bespoke social media analytics software suite, Chorus[3]. This serves to demonstrate the kinds of issues investigators may find themselves contending with, as well as helping figure out ways of handling them methodologically. Our reflexive focus on the research process itself is very much a mainstream methodological practice of social scientists[4] – we seek to take a self-critical view on the (opaque) process of undertaking research involving data collection through the Twitter APIs and data visualisation using spatial mapping algorithms. However, our approach sees such limitations not as *obstacles* to research to be overcome. Rather, we discuss these issues as an exercise in learning the tools of the trade of social media analytics and understanding how they construct the data we analyse, so as to be better able to deal with them as part of our work.

Two Practical Issues

API Rate Limiting in Twitter Data Collection

Twitter data collection is a process mediated through Twitter's various APIs. For the purposes of social media analytics, the APIs are the tools by which users can make requests for specific types of data. This process of using Twitter's APIs to access data necessitates that users write requests as a RESTful statement to return responses in a data interchange format called JSON[5], or that users take advantage of a third-party client which facilitates the task for non-coders. However, though comprehensive collections of data are available for purchase through data archiving services such as Gnip or DataSift, the Twitter APIs are not completely openly available to users and developers. In fact, several restrictions on their usage are in place; ostensibly this is to prevent misuse of the service by individual users. One such restriction is Twitter's 'rate limiting' of an individual account's API usage, or the

rate at which a user can poll the API with requests for data matching a search query. Each API has different rate limits and different software tools will handle those limits in different ways[6]. Providing concrete and quantifiable definitions of these limits is, however, a hopeless task. The 'Search API' (that our example below draws upon) is fairly well-documented in terms of its limits, as we go on to discuss. However, the usage restrictions of all Twitter APIs are variably dependent on contextual information; chiefly the volume of data any queries will yield. Hence, for APIs other than Search, Twitter does not provide information as to the exact limits they will impose. All of this makes understanding and navigating through the Twitter APIs a labyrinthine process. We use data collected via the Search API – which handles data based on keywords appearing in tweets and is the most commonly-used Twitter API – to demonstrate how the API itself inevitably comes to feature in a burgeoning assemblage built up by the research as it is undertaken.

At the time of writing, Twitter's Search API allows Chorus' data collection programme, TweetCatcher Desktop (TCD), to make 450 requests every fifteen minutes. Each such request has the potential to capture a maximum of 100 tweets. The Search API allows users to retrieve historical data up to seven days prior to the initial request. On 23 July 2014, using TCD we ran a very general search query for all usages of the term 'black', as a way of exploring the topics and sub-topics of racialised tweet content. To capture the data, we refreshed the query at various points over an approximately four-hour period so as to ensure as comprehensive a dataset as the API would allow. This resulted in a dataset of 28,317 tweets featuring the term 'black'. Plotting the data over time in half-hour intervals within Chorus' visual analytic programme TweetVis, it was clear that there were gaps in the chronology of the conversation:

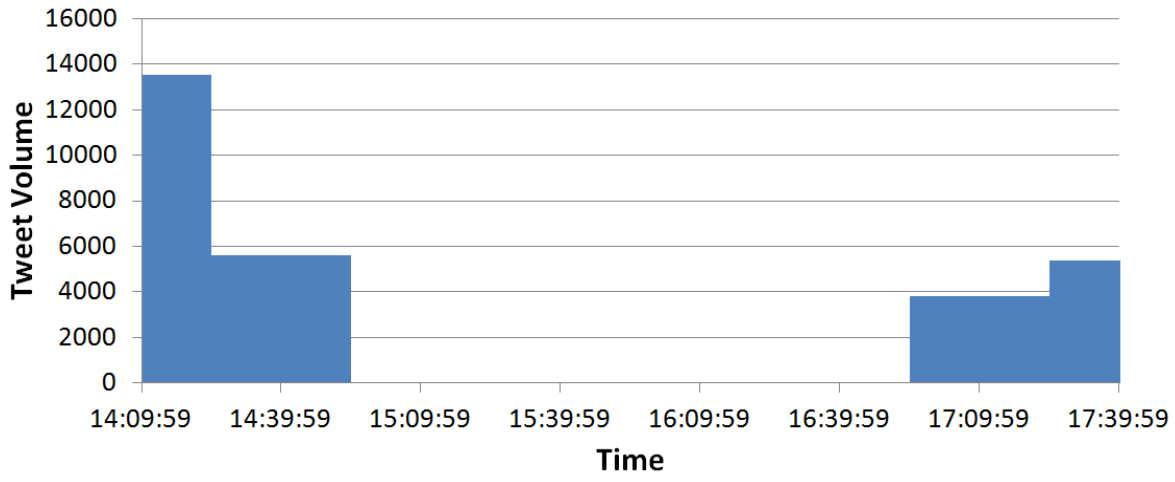

Figure 1: Chart to show volume of tweets mentioning "black" across time (half-hour intervals)

What we see in Figure 1 is a striking reminder that Twitter's APIs are restricted, and that this may have significant effects on the data we wish to capture through them. For high-volume queries it is easy to come up against Twitter's API rate limits, such as during searches for

general terms like 'black', as well as for trending terms (e.g. 'Obama' in the run-up to the 2012 US presidential election). In this example, we were simply unable to keep up with the pace of peoples' tweets; we were able to capture an average 118 tweets-per-minute over the four hour period, whereas the actual conversation skipped along between 450-550 tweets per minute. Naturally, this left a sizeable chunk of data missing from our dataset (see Figure 1). However, what is less immediately obvious in this rendering of the data are the presence of other breaks in the flow of the conversation we captured, which become more apparent when viewing at a finer level of granularity:

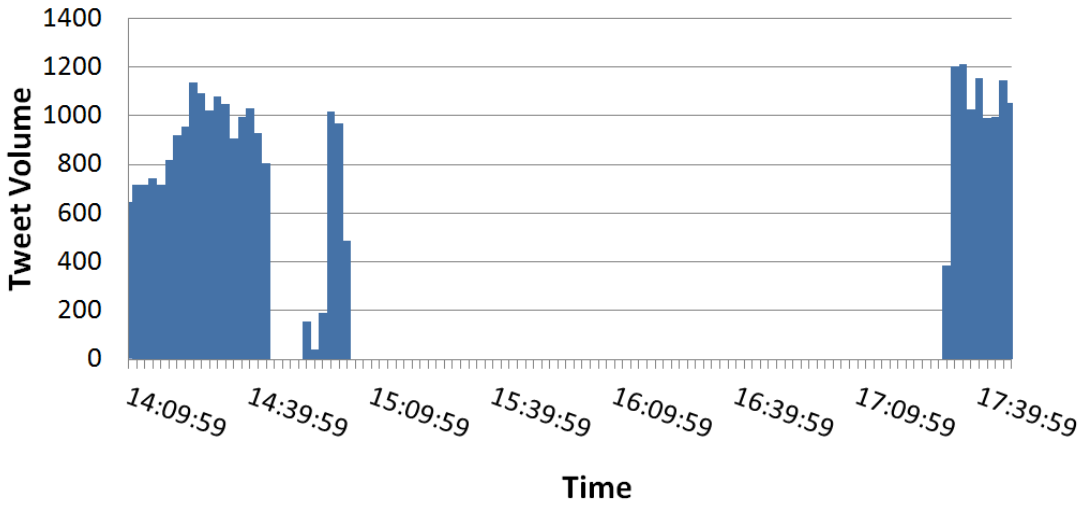

Figure 2: Chart to show volume of tweets mentioning "black" across time (two minute intervals)

In Figure 2, we see the same data grouped into intervals of two minutes. Here we can identify the same gap in the data as in Figure 1, but also an earlier gap which was previously obscured when viewed with our earlier half-hour intervals. Hence, we now can detect a probable disruption to the flow of data between 14:44:59 and 14:58:59 where only a consistent chronology was visible (or at least presumed) before. It may simply be that people tweeted fewer times during these minutes, though it is equally possible that it is at this point we were reaching the limits of what the API would allow us to see. It is in fact impossible to figure out what has happened from the data or visualisations themselves.

What does this ambiguity mean for social media analytics and social research involving Twitter data? A key insight to draw from this demonstration is that comprehensive collections of Twitter data are not freely available to researchers. Even where we may assume we are capturing the entirety of a conversation, drilling down into finer levels of granularity may show us otherwise. Furthermore, failing to recognise when these rate limiting issues have occurred may be detrimental to the analyses we draw from our data. Without due care and attention, we may find ourselves using falsely-derived conceptualisations of data as chronologically consistent[7].

It is important here to acknowledge that access to social media data is a highly politicised issue largely driven by commercial concerns (boyd and Crawford, 2012; Rogers, 2013). In this sense, it is a fallacy to believe that any data which is collected through Twitter's APIs (rather than purchased) is complete: incompleteness and unrepresentativeness are fundamental features *purposefully built into* the APIs to protect the primacy of Twitter's approved data providers. Recognising, understanding and accounting for this is a key step in acknowledging the research process as an unfurling assemblage of interconnected socio-technical entities (of which the API is one, alongside any software and hardware used in the undertaking of the work, the social media users whose posts make up the data, any social theories we use to interpret the resulting findings, and so on). However, the incompleteness and unrepresentativeness of social media data does not prevent us from accessing meaningful insights. It is worth questioning our fetishising of data in this respect – what do we need a chronologically complete dataset *for*? And what can we do without one? Rather than bemoan the purposes for what our data cannot be used, it may be more productive to explore what it *can* do. Though the methodological and analytic possibilities are impossible to encapsulate fully in the present chapter (in that they will depend largely on the questions being addressed), one such approach is advanced in the following section. However, the salient point remains that perhaps the best way to make sense of data is to attain a deep understanding of how a dataset has been constructed, and use that understanding as a resource for designing appropriate analytic approaches with which it may be dealt.

Spatial Mapping Algorithms in Twitter Data Visualisation

Clearly there are issues concerning data collection of which researchers in social media analytics would do well to be aware. However, our endeavours in the field have also revealed similar technical issues in data visualization, where collected data is given an analytic relevance through algorithmic processing. There are already multitudinous tools for visualising social media data – Gephi, NodeXL and Chorus for instance. Some of these utilise spatial mapping algorithms – computational processes through which entities such as individual words or connected arrays of tweeters (or indeed any other kind of 'node') are located on a 2D visual plane in relation to each other. Though each software package operates uniquely, a unifying feature of these algorithms is their use of mathematical reasoning as a way of representing distinctions between nodes. For example, the Chorus visual analytic suite – TweetVis – features (amongst other visualisations) a topical 'cluster map' which uses a spatial mapping algorithm to plot individual words contained within tweets, in regards to the frequency of co-occurrences words have with other words in the dataset. In this map, words which commonly co-occur in tweets cluster together, thereby forming distinct topical clusters through which users can navigate and explore. Here, the algorithm is what constructs and constrains the map, and for users trying to read the visualisation, the constructions and constraints of the map become an integral part of the resulting analysis. We demonstrate the possible effects of this algorithmic constructing and

constraining on analyses, showing how an understanding of the technical goings-on of a data visualisation is a necessary requirement for those wishing to view it sensibly through the lens of an assemblage.

To exemplify what the effects of a spatial mapping algorithm might look like in the undertaking of a social media analytics project, we draw on previous work[8] on 'racialised hashtags' (in particular, the hashtag *#notracist*). With a dataset collecting all usages of the term *#notracist* over an eight-month period (resulting in 24,853 tweets), we plotted a cluster map of hashtags, to see which hashtags featured together more commonly:

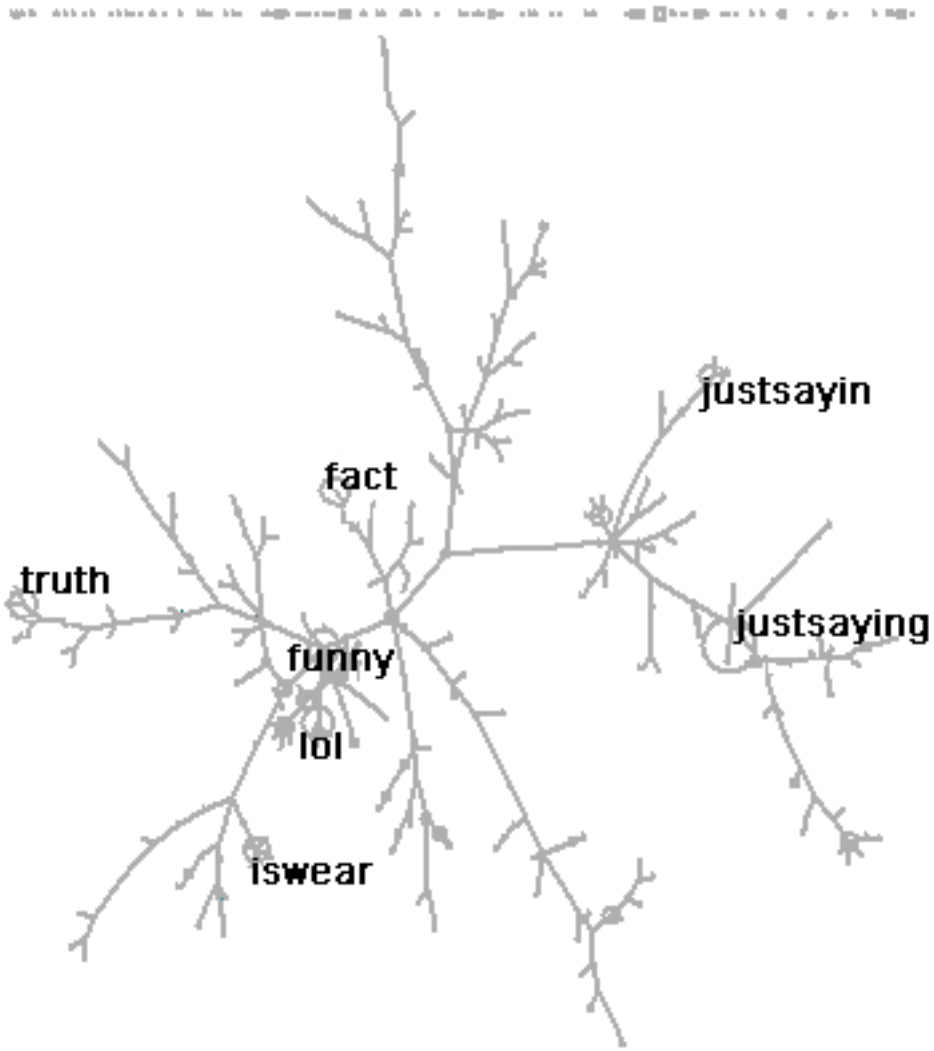

Figure 3 - hashtag map of *#notracist* (labels given to hashtags featuring in =>1% of tweets)

What this map showed us is that there are two distinct types of hashtag being used across the *#notracist* dataset: firstly, a collection of 'comedy' hashtags (including *#funny* and *#lol* and others) located in a tightly-organised cluster near the centre of the map; and secondly, an array of 'truth' hashtags (including *#justsayin*, *#justsaying*, *#iswear*, *#truth*, *#fact* and others) appearing on the fringes of topical branches around the outskirts of the visualisation[9]. Given that these different themes are located in different areas of the map – 'comedy' in a tight central cluster and 'truth' out towards the fringes of branches – we developed an interest in finding out what exactly this difference might be. To drill into the data further, we filtered the master *#notracist* dataset into two sub-sets: one containing 'comedy' hashtags and one containing 'truth' hashtags. We then plotted cluster maps for all terms contained within either dataset:

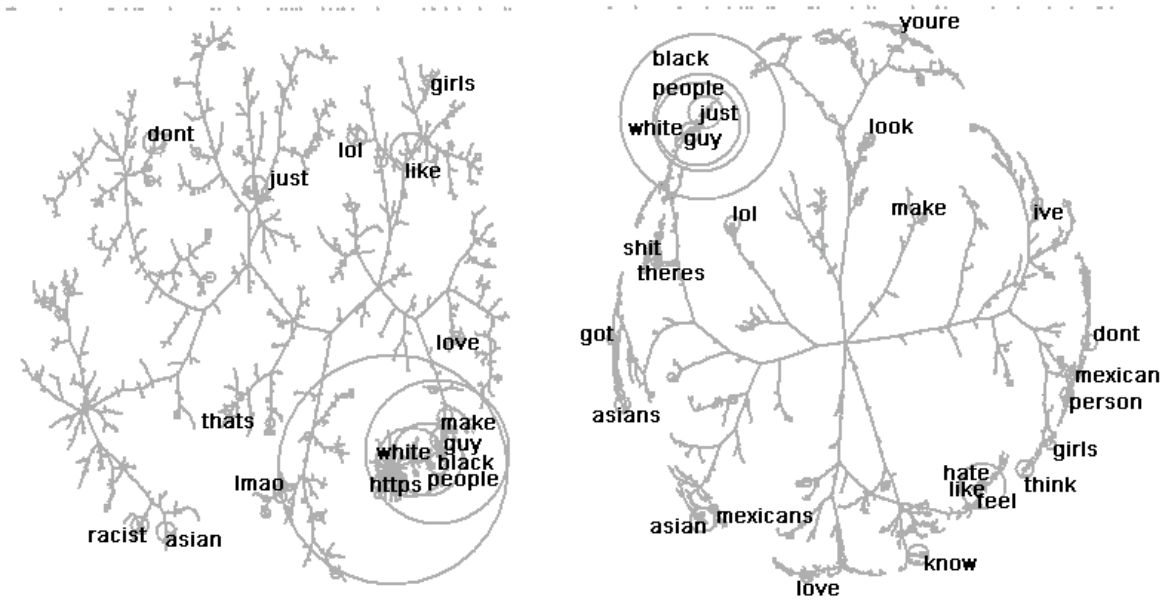

Figure 4 - 'Comedy' term map (left) and 'truth' term map (right) (labels given to terms featuring in =>4% of tweets)

It is clear that these two visualisations are very different from one another – the 'comedy' map is a messy aggregation of highly interconnected terms, whereas terms in the 'truth' map are densely populated around the outer fringes of connecting branches. In order to interpret what the two contrasting maps were telling us, we relied on an understanding of what the algorithm had done with the data points. For the 'comedy' map, the terms used in tweets were highly related to each other showing that *#notracist* comedy was a practice of tweeting done with lots of different terms and hashtags being used in similar ways (i.e. there are a number of 'stock' formats through which tweeters could accountably claim to be 'doing a joke'). However, for the 'truth' map, our understanding of what we could see in the visualisation relied upon an understanding of the functions of Chorus' spatial mapping

algorithms. In this map, terms are chiefly located on the outer edges, as far apart from each other as the algorithm will allow. This is demonstrated by the tree-like appearances of topical branches, with virtually no connecting terms in the centre but a high density of terms pushed out towards the edges of the 2D plane. This visible pushing of the boundaries of the algorithm tells us a lot about the data we were working with. The terms used in the 'truth' tweets we identified are typically disconnected from each other and not used together, and we can begin to characterise *#notracist* 'truth' tweets as evidence of a topic that is not implicitly agreed-upon and which reflects a diverse array of strategies for justifying a *#notracist* claim as a statement.

Again, we ask: what does this mean for social media analytics? Visualisations such as those discussed here are designed to fit data into a model which serves to constrain our data and analytic materials as well as give them visible structure[10]. Hence, our aim in describing the processes through which such models construct and constrain analyses is to set out a requirement for social media analytics that it explicitly *account for* these processes. For example, in the *#notracist* data described above, it was fundamental to our understanding of the data that we could recognise that 'truth' terms were pushing the Chorus cluster map algorithm to its limits, and consequently use that information as a way of figuring out what the 'truth' conversation was about. Crucially, the fact of our data being processed through the algorithm in a certain way is precisely what helped us get to grips with what lay at the heart of that data. Without this we would have been lost. Consequently we reiterate that these intrusions of computational, technical and mathematical processes into our analyses is not something to be resisted. Rather, they are necessary and productive elements of social media analytics without which we would be unable to characterise the assemblage through which the analysis had been shaped, and ultimately, unable to make adequate sense of the data at hand.

Concluding Remarks

In this chapter, we have outlined how social media analytics incorporates data collection and analytic work in ways which are thoroughly reliant on computational and technical processes. Despite the provocative nature in which the question in the title of this piece was posed, we believe investigators in the field *are* in a position to sensibly account for their work in terms of these processes. Yet these issues are not something we have seen discussed *explicitly* in the accounts produced thus far. Hence, though our points may be frustratingly mundane to our peers (who may question why researchers would want to write about the inner workings of APIs and algorithms) we nonetheless think it valuable to discuss such things transparently as a means of promoting healthy and robust methodologies for the emerging field.

To that end, rather than depict software tools in general and in abstract, we have exemplified our ideas with reference to two specific issues arising out of the use of just one

software package. This, we hope, gives a flavour of *the kinds of* issues of which researchers must be aware when working with digital data and associated software tools. It is our hope that our accounts of two specific examples can demonstrate just how these kinds of issues intersect with research work in a very direct way. Working in this way, we have shown how technical and computational processes become a 'necessary evil' of the work. Only there is nothing 'evil' about them. Rather, these same processes can be used as *resources* for conducting (and figuring out how to conduct) analytic work in appropriate ways. However, the use of these processes as analytic resources relies on our having a deep understanding of what they are doing with our data, else we risk wrong-footing our analyses before we even begin. As Manovich notes:

> ...you must have the right background to make use of this data. The [analytic] software itself is free and readily available...but you need the right training...and prior practical experience to get meaningful results (Manovich, 2012: 472).

It is clear that it now becomes our job as researchers to equip ourselves with these understandings of the technical processes on which our work relies, however much this may take us outside of typical disciplinary boundaries.

All of this may make the analysis of social media data an infinitely more complex issue, in that we are no longer really analysing *only* the data, but an assemblage (Langlois, 2011; Sharma, 2013) of technical and social processes which coalesce to form the datasets and visualisations we find before us. Concerning data collection, we have used the idea of an assemblage to outline how the technical aspects of API rate limiting become built into social media analytics research from the very beginning of the research process. Concerning data visualisation and analysis, our described assemblage relied upon our conceptualisation of computational processes as having (by necessity) a commitment to numerical 'understandings' of data and how those 'understandings' are translated into images to be read by human researchers. We have no doubt that Chorus' way of doing things is only one amongst many, and other such issues will invariably arise in a multitude of different ways when doing social media analytics with other tools. As with any software tool, Chorus is not 'just a tool' – it engenders a particular way of thinking about social media data which constructs and constrains analyses in equal amounts. As such, the modest goal of this chapter has been to encourage readers to consider their research work not only in terms of the results and findings to be drawn, but in relation to the myriad processes through which those findings are mediated throughout the endeavour. We advocate *thinking in assemblages* as a requirement for social media analytics generally. Furthermore, we have shown the relevance of assemblages for mainstream purposes, and how the specific properties of an assemblage might be uncovered through the deployment of a key methodological principle – reflexivity – which has informed the present chapter from start

to finish. In using the idea of assemblages as a frame for undertaking investigative work, analytic findings would be explicitly situated alongside deconstructions of the processes by which tools are governed by big data and the processes by which those same tools govern the generation of empirical findings. In this sense, we may find ourselves in the business of handling *data processes* rather than data, and of reading *visualisation processes* rather than visualisations. The final result of these processes – the compiled dataset or the visual representation – are not objects in and of themselves, but are better thought of as a way of demonstrating how an unfolding combination of human and computational research processes has resulted in a selection of valid and defensible findings.

Bibliography

boyd, d. and K. Crawford (2012) 'Critical Questions for Big Data', *Information, Communication & Society*, 15(5), 662-79.

Fielding, N. and R. Lee (2008) 'Qualitative e-Social Science/Cyber-Research' in N. Fielding, R. Lee and G. Blank (eds.) *The Sage Handbook of Online Research Methods*. London: Sage, pp. 491-506.

Hine, C. (2006) 'Virtual Methods and the Sociology of Cyber-Social-Scientific Knowledge' in C. Hine (ed.) *Virtual Methods: Issues in Social Research on the Internet*. Oxford: Berg, 1-13.

Keegan, B. (2013) 'C-Level Executives Cry Out for Data Scientists', *ComputerWeekly.com*, [ONLINE] Available at: http://www.computerweekly.com/news/2240205984/C-level-executives-cry-out-for-data-scientists [Accessed 13 November 2014].

Langlois, G. (2011) 'Meaning, Semiotechnologies and Participatory Media', *Culture Machine*, 12, 1-27.

Lewis, S. C., R. Zamith & A. Hermida (2013) 'Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods', *Journal of Broadcasting & Electronic Media*, 57(1), 34-52.

Lynch, M. (2000) 'Against Reflexivity as an Academic Virtue and Source of Privileged Knowledge', *Theory, Culture & Society*, 17(3), 26-54.

Kitchin, R. (2014) 'Big Data, New Epistemologies and Paradigm Shifts', *Big Data and Society*, 1, 1-12.

Manovich, L. (2012) 'Trending: The Promises and Challenges of Big Social Data', in M. K. Gold (ed.) *Debates in the Digital Humanities*. London: University of Minnesota Press, 460-75.

Marres, N. (2012) 'The Redistribution of Methods: On Intervention in Digital Social Research, Broadly Conceived', *The Sociological Review*, 60(S1), 139-65.

Murthy, D. (2013) *Twitter*. Cambridge: Polity Press.

O'Connor, H., C. Madge, R. Shaw and J. Wellens (2008) 'Internet-Based Interviewing' in N. Fielding, R. Lee and G. Blank (eds.) *The Sage Handbook of Online Research Methods*. London: Sage, 271-89.

Procter, R., F. Vis and A. Voss (2013) 'Reading the Riots on Twitter: Methodological Innovation for the Analysis of Big Data', *International Journal of Social Science Research Methodology*, 16(3), 197-214.

Rogers, R. (2013) *Digital Methods*. London: The MIT Press.

Sharma, S. (2013) 'Black Twitter? Racial Hashtags, Networks and Contagion', *New Formations*, 78, 46-64.

Tufekci, Z. (2014) 'Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls', *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, AAAI Publications, 505-14.

[1] We take 'big' social media data to refer to volumes of data too large to handle without computational processing and which are derived from peoples' everyday usages of social media platforms such as Twitter.

[2] The idea that social media analytics requires no specific skill in its practitioners is contestable – for instance, Keegan (2013) notes that the information technology industry believes itself to be suffering from a lack of trained data scientists. However, the point remains that there are lots of freely available social media analytics tools with which investigators from any discipline can explore data, and it is no longer a steadfast requirement for practitioners to have any significant skills in programming, data mining, data visualisation, and so on.

[3] Chorus (see www.chorusanalytics.co.uk for further details) is a free-to-download data collection and visual analytic software suite dealing with Twitter data for social media analytics. Chorus was developed at Brunel University by a team including several of the authors of the present chapter (Dr. Tim Cribbin, Dr. Phillip Brooker and Prof. Julie Barnett). The development of Chorus was supported in part through the MATCH Programme (UK Engineering and Physical Sciences Research Council grants numbers GR/S29874/01, EP/F063822/1 and EP/G012393/1).

[4] Lynch (2000) however questions the utility of sociology's concern with self-analysis, arguing that it only need be applied when something particularly interesting will result (as is the case with his own reflexive approach to reflexivity, and we hope, the present chapter).

[5] A RESTful statement is one which is written in adherence with REST (or Representational State Transfer) principles, REST being the ubiquitous architectural style that standardises and underlies the world wide web. In regard specifically to handling the Twitter APIs, RESTful statements are the commands by which API users can speak to Twitter's servers to request specific slices of data, which are returned in JSON format. Readers wishing to find out more about using the Twitter APIs and writing API requests manually should start with Twitter's own developer documentation, available at: https://dev.twitter.com/docs [accessed: 29 July 2014].

[6] See https://dev.twitter.com/rest/public/rate-limiting [accessed: 17 November 2014] for a more detailed account of the rate limiting Twitter applies to its APIs.

[7] Other APIs may provide something more like a chronologically complete timeline – for instance, the Twitter Streaming API pushes 'real-time' data matching a query's criteria. However, the Streaming API only provides a percentage sample of tweets requested, where the actual percentage is unknowable and dependent on the volume of tweets requested by the query and concurrent Twitter traffic. Hence, the only way to ensure comprehensivity of a dataset without running into sampling issues is to purchase Twitter 'Firehose' data – this alone politicises access to data to the extent that few can afford to ever see a comprehensive dataset.

[8] See http://www.darkmatter101.org/wiki/notracist_twitter [accessed: 21 January 2015] for an informal account of this project.

[9] Though this chapter is not intended as an empirical study of these tweets, interested readers might wish to note that the 'comedy' hashtags we identified were tweets designed by tweeters to be humorous, whereas 'truth' hashtags were designed to enforce a point that a tweets was 'just a fact' or 'just an observation' and so on. Our analytic work explored the different practices through which users attempted to justify their claims that a tweet was not racist by virtue of it being a joke or a statement of fact.

[10] This might be likened to filling a glass with water. As with water taking the shape of the glass it is poured into, the process of collecting and visualising social media data serves to furnish amorphous data with a structure. However, as much as these technical processes *construct* data such that they become amenable to analysis, it can be said that the same processes *constrain* data into singular readings – a circular glass gives only a circular shape to the water, but what if other shapes would prove more interesting?