View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Bath Research Portal

Citation for published version:
Brooker, P, Barnett, J & Cribbin, T 2016, 'Doing Social Media Analytics', Big Data & Society, vol. 3, no. 2, pp. 1-12. https://doi.org/10.1177/2053951716658060

DOI:
10.1177/2053951716658060

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Link to publication

Publisher Rights
CC BY

Brooker, P. Barnett, J. Cribbin, T., Doing Social Media analysis, Big Data & Society (Vol: 3 Iss: 2). Copyright © 2016 The Author(s). Reprinted by permission of SAGE Publications.

University of Bath

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 13. May. 2019

Original Research Article

Big Data & Society
July–December 2016: 1–12
© The Author(s) 2016
DOI: 10.1177/2053951716658060
bds.sagepub.com

Doing social media analytics

Phillip Brooker[1], Julie Barnett[1] and Timothy Cribbin[2]

Abstract

In the few years since the advent of 'Big Data' research, social media analytics has begun to accumulate studies drawing on social media as a resource and tool for research work. Yet, there has been relatively little attention paid to the development of methodologies for handling this kind of data. The few works that exist in this area often reflect upon the implications of 'grand' social science methodological concepts for new social media research (i.e. they focus on general issues such as sampling, data validity, ethics, etc.). By contrast, we advance an abductively oriented methodological suite designed to explore the construction of phenomena played out through social media. To do this, we use a software tool – Chorus – to illustrate a visual analytic approach to data. Informed by visual analytic principles, we posit a two-by-two methodological model of social media analytics, combining two data collection strategies with two analytic modes. We go on to demonstrate each of these four approaches 'in action', to help clarify how and why they might be used to address various research questions.

Keywords

Social media, Twitter, analytics, digital social science, data visualisation, methods

The ever-expanding usage of social media throughout everyday life offers a critical data resource to social scientists. Though this is increasingly recognised, such data brings with it new methodological challenges in terms of finding ways to analyse what it tells us about social life. Social media provides a form of user-generated data which may be unsolicited and unscripted, and which is often expressed multi-modally (i.e. through combinations of text, hyperlinks, images, videos, music, etc.). Hence, it is important to consider the challenges that this data holds for researchers in terms of rendering them amenable to analysis, and in identifying the sort of research questions that such data might appropriately address. As Raghavan (2014) notes, researchers no longer lack computational tools or theories to help make sense of social media data, yet there remains a paucity of methodologies to make transparent the move from tools to explanations.

We address this challenge by demonstrating the value of a 'visual analytic' approach to capturing and exploring the qualitative and subjective facets of Twitter data as a socio-technical research 'assemblage' (Langlois, 2011; Sharma, 2013) wherein the phenomena uncovered by research are acknowledged as essentially intertwined with the technical aspects of data collection and visualisation (amongst other aspects of the research process more generally). We choose Twitter as a foundation due to its role as 'a ''model organism'' of big data' (Tufekci, 2014: 506). Twitter is widely used[1] and sufficiently simple in its broadcast mechanics such that its exploration 'is conducive to progress in basic questions underlying the entire field' (Tufekci, 2014: 506).[2] In addition to 140-character-limited linguistic content, tweets are accompanied by various metadata, including: timestamps; tweeters' usernames and userIDs; 'follower' and 'following' counts; geo-location coordinates; hashtags; @mentions (i.e. communications between users); retweets (where users re-post others' tweets) and hyperlinks. Coupling the lexical

[1]University of Bath, UK
[2]Brunel University, UK

Corresponding author:
Phillip Brooker, University of Bath, 3rd Floor, 10 West, University of Bath, Claverton Down, BA2 7AY, UK.
Email: p.d.brooker@bath.ac.uk

Creative Commons CC-BY: This article is distributed under the terms of the Creative Commons Attribution 3.0 License (http://www.creativecommons.org/licenses/by/3.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Downloaded from by guest on July 18, 2016

content of tweets with these metadata provides a rich context in which to base an analysis.

The paper is organised as follows. We first discuss existing methodological literature in social media analytics, highlighting the shortage of methodological strategies for handling social media data. We go on to posit a (visual analytic) framework that seeks to address this, grounded in an abductive ontological perspective. We outline the general ideas behind two modes of data collection and two analytic approaches, exploring their four combinatory permutations. We demonstrate the four analytic lenses with empirical examples: the role of media in talk around the UK 2011 e coli food scare, user experiences of adrenaline auto-injectors ('epipens'), bovine tuberculosis (bTB) and UK badger activism, and symptom-reporting amongst cystic fibrosis sufferers.

Existing social science approaches to Twitter

Many studies of Twitter focus on theoretical/conceptual issues (e.g., boyd & Crawford, 2012; Kitchin, 2014; Matthews and Sunderland, 2013; Murthy, 2012, 2013) or present empirical studies (e.g., boyd et al., 2010; Burnap et al., 2014; Heverin and Zach, 2011; Tumasjan et al., 2010). Methodology, the third facet of this triumvirate, is only just beginning to receive the same attention.

Many important methodological contributions (e.g., Brügger and Finnemann, 2013; Mahrt and Scharkow, 2013; Matthews and Sunderland, 2013; Rogers, 2013; Tufekci, 2014) begin with grand methodological concerns long standing in the social sciences such as sampling and demographics, and the representativeness and quality of social media data. For instance, Mahrt and Scharkow (2013) contend that 'it is currently impossible to collect a sample in a way that adheres to the conventions of sample quality established in the social sciences' (p.22) and Rogers (2013) notes:

> To view the web as data set for social and cultural research is to be confronted with a variety of issues about messy data... Here the general reputation problem about quality online is transformed, initially, into the question of how to clean up the data, since there is a lack of uniformity in how users fill in forms, fields, boxes and other text entry spaces. (Rogers, 2013: 205)

Here, Mahrt and Scharkow and Rogers question the capacity of social media data to capitulate to an interrogation via methods we already have at our disposal. This application of existing methodological thinking to social media analytics gives us pause – to what extent are these concepts relevant to social media analytics, given that its difference from more traditional forms of enquiry is often considered a defining characteristic? We build on previous work (Anderson, 2011; boyd, 2010; Bruns and Burgess, 2011; Gillespie, 2014) to explore methodologies which are more attuned to the newness inherent in social media analytics.

As such, we align our work less with authors dealing with the trends and patterns available in aggregate overviews of large-scale data, and more with Marshall (2012), who provocatively destabilises the processes through which tweets come to be construed as data, advocating a more qualitative-flavoured approach based on closer readings of tweets. In this regard, our approach downplays the relevancy of concepts such as data validity and sampling as yardsticks for measuring the 'goodness' of social media data. Instead, we explore what is possible with the 'noisy' and 'unclean' data we have to hand, providing an alternative means to 'drill down' into the substantive content of tweets.

The proposed methodological toolkit is supportive of abductive modes of enquiry (Blaikie, 2000; Locke, 2010), which work through a research process inductively, towards increasingly plausible explanations of phenomena. Ontologically, abductive reasoning operates with an understanding of research phenomena as co-constructed products of social interactions, including those which contribute towards the undertaking of research by researchers. Such phenomena do not exist outside of the 'assemblage' of social, technical, material and other factors which converge to generate, shape and make them available to social research. The idea of assemblages is implicated in the work of boyd (2010), Bruns and Burgess (2011) and Gillespie (2014). Here, the interactions of a digital public – a 'community' of people brought together in and by a digital domain – are shaped by the affordances of the platforms they use (i.e. a social media service). This shaping works both ways: as a crude example, Twitter's conventions help structure how people communicate and what they use the platform to say, and people's tweeting practices serve to reorganise the possibilities of Twitter (e.g., the oft-remarked-upon 'folk' origins of the hashtag). In practice, investigating assemblages necessitates the incorporation of a wider, more inter-connected array of elements, including: users, platforms, communicative practices, cultural events and issues (about which people communicate), the algorithms that structure their interactions, the research process itself (as the mechanism by which a researcher develops insights), and so on.

If our subjects are the messy products of such interweaving factors then, as Hughes and Sharrock (1997) note, epistemologically we must (unapologetically) concede that our descriptions of the reality at hand

Downloaded from by guest on July 18, 2016

inevitably describe only the assemblage we have built to invoke that reality's construction. Accordingly, we direct our methodological toolkit towards supporting the task of making transparent *how* these assembled factors produce research findings. Under this schema, we begin to see and interrogate the processes through which users, algorithms, offline 'real-world' occurrences and other social factors collectively emerge as tools of public knowledge and discourse (Anderson, 2011). To this end, we advance a 'visual analytic' approach to facilitate the unpicking of Twitter and the investigative process itself as a socio-technical assemblage (Langlois, 2011; Sharma, 2013).

A visual analytic framework for understanding Twitter data

Visual analytics (Thomas and Cook, 2005) integrates techniques from two fields – information visualisation (Card et al., 1999) and computational modelling – and has already made notable contributions to social media analytics (Diakopoulos et al., 2010; Hassan et al., 2014). The key principle of visual analytics is the use of interactive visualisation as part of a process of analytical reasoning (rather than as static outputs displaying the outcomes of an analysis) (Thomas and Cook, 2005). Central to the approach is that visualisations do not replace the skills of the researcher but rather amplify their inherent capabilities by capitalising on the high-bandwidth processing of visual perception (Card et al., 1999). The analytic process begins with overviews (high-level abstractions) that guide researchers towards potentially interesting aspects of the data. The researcher can then either transform the view (e.g., by zooming or filtering) or create new visualisations that enable them to 'drill down' and engage in finer qualitative analysis of their data.

This visual analytic approach to social media data is a natural extension of abductive interpretivist reasoning, in that it promotes a thoroughly exploratory orientation to data and analysis. Reflecting the ontological and epistemological ideas outlined above, as new insights are derived they are fed back into the assemblage and the research process becomes an endlessly exploratory endeavour – such exploratory work is not merely prior to 'a proper analysis'; it *is* the analysis. With data collection and analysis as one seamless process, exploration of an initial dataset may lead to new questions which may in turn feed back into new rounds of data collection, with findings emerging throughout the ongoing iterations. Hence, visual analytics facilitates a mind-set wherein researchers can probe their own assumptions and perspectives to help capture phenomena as they unfold and encourages questioning around the relevance of long-held social science concepts to evaluate their applicability to new digital data (as we have attempted to do above).

The framework and examples we present here apply visual analytics to social research questions. We have undertaken our empirical work using the Chorus data collection and visualisation suite,[3] which provides a set of affordances to researchers via two types of visualisation: the timeline and cluster map views outlined in the examples below. Our usage of Chorus' visualisations is not undertaken with the aim of concretising a singular reading of the data – we reiterate that the visualisations do not demonstrate singular readings, and to read them as if they were is to downplay the analytic impact of the various processes through which they come to be. In this regard, the analytic strategies we outline below demonstrate ways of using visualisations to navigate around social media data and work towards qualitative insights grounded in original tweet content.

Commensurate with our adopted abductive ontology and the notion of the research process as an assemblage, we acknowledge that our methodological talk is inextricably bound to the affordances Chorus provides. This need not disrupt the sense in which the present paper can speak to the methodological exigencies of social media analytics research, since we are not positing a set of abstract techniques for handling social media data. No such taxonomy of techniques is possible, and every new research project will demand different techniques specifically tailored to the data assemblage at hand. We instead aim to demonstrate and facilitate a methodological mind-set of 'thinking in assemblages', which will help researchers generate techniques of their own to suit their work.

'Thinking in assemblages' (via a visual analytic approach) offers researchers a way to be reflexively attuned to their phenomena and to the role of research as a subjective enterprise. This is vital for qualitative research in that it ensures researchers are better able to account for their phenomena by demonstrating the processes through which findings are derived. This endeavour is tied to the application of the visual analytic principle of folding visualisations into the research process (as opposed to treating them as results), and in the understanding of research as iterative. It is on these two points where we hope to extend our approach to research conducted with other tools (and also to social media platforms other than Twitter).With this aim in mind, we posit a framework which outlines a set of choices for researchers to make in terms of how to collect and analyse their data. This framework codifies several complementary research approaches to Twitter research; rather than pitch these as 'novel', our aim is to give methodological shape to social media analytics by situating various existing approaches in relation to one another. As part of this framework, we outline two

Downloaded from by guest on July 18, 2016

data collection strategies (*semantically driven* and *user driven*) and two analytic modes (*temporal analysis* and *corpus analysis*), as different ways of organising Twitter data. We provide empirical examples for the four resulting permutations, indicating the sorts of research questions they might be used to address and the kinds of insights they might uncover.

Data collection

Twitter's APIs (Application Programming Interfaces; the technologies through which users access Twitter data) allow users to retrieve a range of data entities and associated values. We outline two approaches to collecting this data. First, the familiar query keyword search, which utilises linguistic entities (i.e. words, hashtags, URLs) as criteria for compiling datasets. Second, we discuss data consisting of extended timelines of groups of users – a user-following strategy.[4]

Capturing semantically driven data (query keyword searches)

This type of data capture takes the semantic content of users' tweets as its starting point. The research process might therefore begin by identifying keywords that are likely to typify tweets around a topic of interest, using logical operators to define the scope. The resulting data has an inherent semantic orientation around a topic, whilst retaining a degree of flexibility as to how exclusive the query is (i.e. it can include a selection of alternative terms to account for variations in the ways people tweet around the topic).

Capturing user-driven data (user following)

User-driven data is organised around the Twitter activity of selected groups of users. This involves identifying users whose tweets are pertinent to a research question, pulling their Twitter timelines and sifting for research-relevant themes. This approach is useful for projects where a keyword query is not easily defined (i.e. where tweeters use implicit, informal, colloquial or general references to the area of interest) or where there is value in understanding the role of a particular issue within a broader set of preoccupations. Whilst allowing researchers to *find out* what a group of people are tweeting about without narrowing the scope with keywords, this strategy nonetheless provides an analytic challenge in terms of the diversity of topics captured.

Data analysis

Complementary to these data collection strategies, we outline two analytic orientations to Twitter data: temporal and corpus analysis. There has been a recent trend in the application of visual analytics towards representing how topical structure evolves over time – see Cui et al., 2011; Luo et al., 2012; Marcus et al., 2011; Rose et al., 2009. However, we propose that there is value in decomposing temporal and semantic structures into distinct but coordinated views of the same data. On the conceptual level a time-dependent event-based view of data and a non-time-dependent topic-based view of data can be conceived as two sides of the same coin that is topic evolution. We do this because there are interesting social science research questions about topics which may not require an insight into how a topic has evolved – see for instance our examples on user experiences of epi-pen devices and symptom reporting by cystic fibrosis sufferers. This, we argue, allows researchers to more straightforwardly see the possibilities of each analytic type before considering how best to combine them. Moreover, our distinguishing between these two approaches does not preclude researchers from exploring topic evolution in the move from one methodological strategy to another – in fact, we encourage this as part of the iterative nature of visual analytics as a social science methodology.

Given our concern to display how software tools become embedded within the assemblages we construct to render social phenomena visible, it is worth noting several technical differences between Chorus and other (aforementioned) tools and approaches. First, in contrast to more general text analytic tools such as Textflow (Cui et al., 2011), Chorus is specifically designed to be sensitive to the technical and contextual exigencies of Twitter, affording a deeper exploration of Twitter's role in the assemblages we build around it. Second, Chorus uses Twitter's Search API rather than its Streaming API (as is the case with TwitInfo (Marcus et al., 2011)) allowing for more comprehensive recall of data around specific topics. Third, the exploration of tweets and user timelines with Chorus' particular spatial-semantic (cluster) views facilitate unique analytic possibilities not provided by other Twitter analytic tools. These features (and more) situate Chorus as a useful alternative to existing tools, the possibilities of which are demonstrated in the example cases below.

Temporal analysis (event based)

Twitter data can be viewed as a temporally unfolding narrative. Researchers may draw insights from such things as: variation in tweet volume around loci, evolving positive or negative sentiment over the course of a conversation, shifts in the vocabulary characterising a discussion, changes in the likelihood of URLs being

Downloaded from by guest on July 18, 2016

referenced within tweets, and so on. In this way, a chronological viewing lends itself to the exploration of 'events' as they unfold within Twitter.

Corpus analysis (topic based)

By contrast, a corpus analysis relies on a conception of whole datasets as an 'information space' in which semantic features (words, hashtags, etc.) intersect in potentially interesting ways, irrespective of the time they are expressed. Researchers may draw insights from the exploration of topical structures emerging from the entire body of data, investigating the ways in which keywords are used together to form broader themes. In this way, a corpus analysis viewing of the Twitter data lends itself to the exploration of 'topics'.

Four empirical examples

Based on the four quadrants of our framework (Table 1) we demonstrate how each can be used to address distinct types of social research questions. These examples should not be considered comprehensive treatments of the data – indeed, a comprehensive treatment is impossible if the social media analytics research process is understood as ever-exploratory and essentially iterative. Rather, our aim is to *point the way towards* a fuller analysis by demonstrating how to collect data that may speak to a particular research question and techniques for analytically handling that data with Chorus' visual models.

Across each of our examples, we outline our data collection strategy, the keywords used or the user timelines selected. Chorus' queries tend towards inclusivity (e.g. a query keyword of 'epipen' would also capture tweets containing the terms 'epipens' and '#epipen'). Queries return all tweets with unique TweetIDs (including retweets), removing duplicate entries where they satisfy more than one query criteria. For the examples below, retweets were not removed.

We also outline our usage of Chorus' visualisations to find our way around the data, and it is helpful here to briefly describe how Chorus builds those visualisations since this is a formative factor in the data assemblages they help create. Chorus first builds a 'word' index containing counts of all significant corpus words within each tweet in the dataset. Less significant words such as stop words ('a', 'the', 'and', and so on), particularly rare or common terms are pruned from the index prior to analysis. By default, words that occur in more than 50% of all tweets or fewer than 0.1% or two tweets (whichever is greater) are removed. This indexing results in a matrix of word-tweet counts from which measures of both tweet and term similarity are computed (using cosine or the normalised dot product metric). Chorus also derives a word-interval matrix, an aggregated version of the term-tweet matrix which contains the standardised (0–1) frequencies for each term in each specified time interval (seconds, minutes, etc.). This is used to compute various temporal statistics described below.

The timeline graphs (Figures 1 and 5) display various statistics including: tweet volume, ratio of tweets containing a URL, positive sentiment, negative sentiment, novelty of terms and homogeneity of terms (see below for further detail). The cluster map visualisations (Figures 2, 3, 4 and 6) use the word index to compute a map in which the distance between words is inversely proportional to their contextual similarity, i.e. words that tend to commonly occur together in tweets are positioned closer together. In this way, groups of related words cohere into 'clusters', providing a thematic overview and a basis for navigation around the dataset.

Temporal analysis of semantically driven data

A **temporal** (or event-based) view of **semantically driven** Twitter data draws on the chronology available both in absolute terms of the time of posting (CreatedAt field) and as a result of relative tweet order (TweetID field).

Table 1. Combinations of different strategies for data collection and analysis.

		Data analysis	
		Temporal analysis (event based)	**Corpus analysis (topic based)**
Data capture	**Semantically driven (query keyword)**	How does a narrative about a semantic entity (i.e. word, hashtags, etc.) unfold over time?	How is talk around a semantic entity organised topically (and sub-topically)?
	User driven (User following)	How do users' language and tweeting practices change (or not) over time?	What topics are a specific group of users tweeting about (and how are they doing it)?

Downloaded from by guest on July 18, 2016

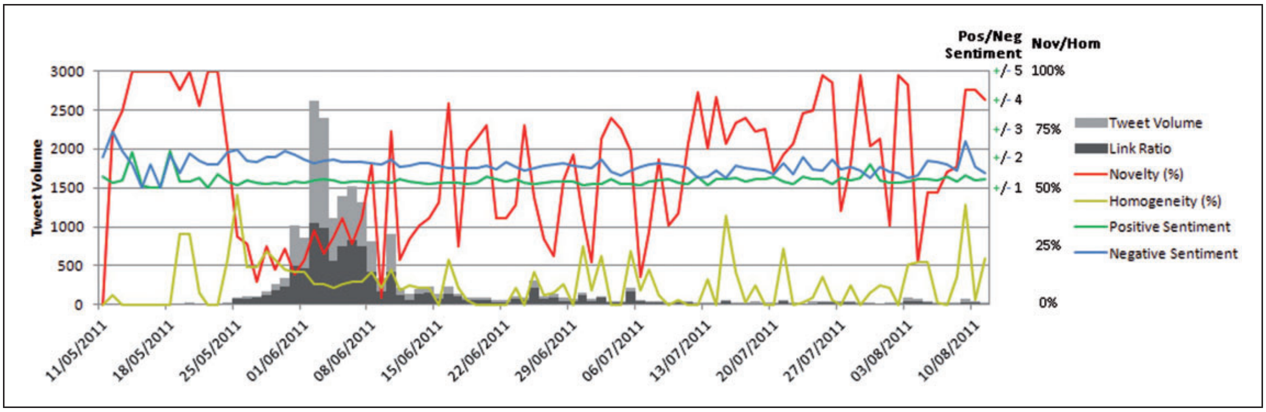

Figure 1. Timeline view of ‘e coli’ data.

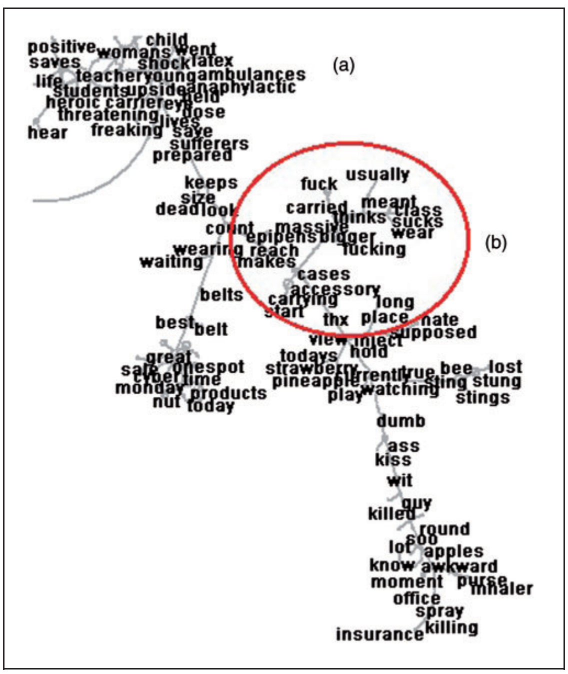

Figure 2. ‘Devices and user-experience’ branch (origin nodes labelled a, physical device user opinions sub-branch labelled b).

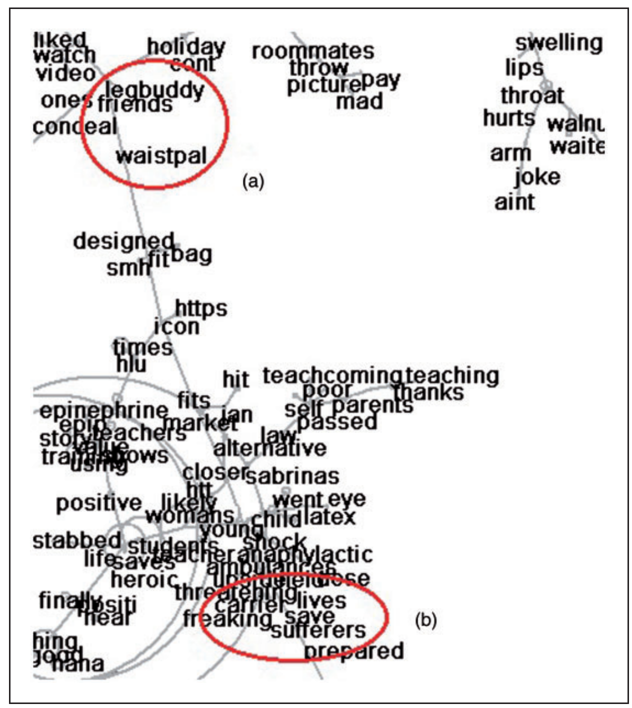

Figure 3. Origin nodes of the ‘devices and user-experience’ branch (labelled a). Note the terms ‘legbuddy→waistpal’ at the root, and ‘carrier→sufferers’ (indicating the relevant user-group (labelled b)).

The semantically driven nature of the data centres the analysis on specified unifying aspects of conversation – a hashtag, mentions to a particular user account, etc. – and associated attributes such as tweet volume, the ratio of tweets to tweets with links, sentiment analytics, semantic homogeneity or novelty, and so on.[5] The focus here is on how information within various data fields fluctuates or maintains across time, providing insight into how people use Twitter across a Twitter-reported event.

Example: The role of media in talk around the UK 2011 e coli food scare. We explored public perceptions of *E. coli* during the 2011 EHEC/*E. coli* bacteria outbreak in Europe. Our search terms were ‘e coli’ and related terms (‘e. coli’, ‘#ecoli’), capturing 19,998 tweets spanning an approximately three-month period (mid-May to mid-August 2011). Each interval in Figure 1 represents a day of tweets. Our interest was in exploring the different periods constituting the Twitter e coli scare ‘event’, characterised by different styles of tweeting within those periods (e.g. ‘fact-sharing’, ‘rumour propagation’, ‘raising awareness’). This time-dependent

Downloaded from by guest on July 18, 2016

view lends itself to research questions concerning change in opinions and meanings: it is *event based*, in that the characteristic narratives of an 'event' are revealed through an unfolding chronological order of key moments.

We identify several stages to the conversation.[6] First, the period from 11 May 2011 to 23 May 2011 is a *precursor to the main event*; it contains few mentions of 'e coli': unrelated jokes, small-scale local news stories and so on. This incoherence is visible in the high level of novelty – the red measure – which indicates that this period contains few terms persisting in an ongoing conversation.[7] A second phase is identifiable at 24 May 2011, with a marked decrease in the semantic novelty of tweets. This is the beginning of a six-day period, 24 May 2011 to 30 May 2011, based around *factual information propagation* and the sharing of news headlines and URLs to news websites. At this point, retweets from news media stories enter into the conversation – links to articles citing the origin of the outbreak in Germany form a significant portion of the total volume of tweets, with between 76 and 90% of tweets featuring a URL. The next period, from 30 May 2011 to 10 June 2011, sees a huge increase in the volume of tweets, whilst the ratio of tweets to tweets with URLs drops considerably and we find people using a different set of terms to talk about e coli (exemplified by the rising novelty metric and the falling homogeneity metric). Tweets here are emotive in content rather than factual, and it is at this point that e coli begins to become a concern for a significantly larger Twitter population who express their anxieties, ask for advice, show sympathy for sufferers and fatalities, and so on.

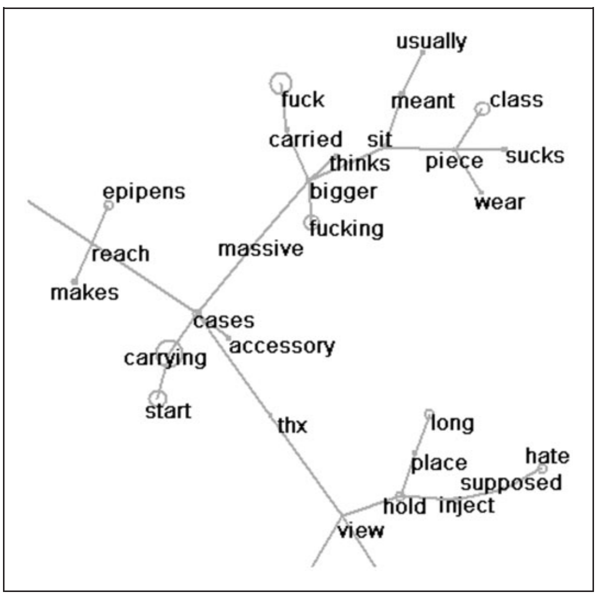

Figure 4. Sub-branch showing user opinions on physical aspects of epipens (labelled b in Figure 2).

With this analysis, we have begun to characterise a Twitter event by breaking it down into time-dependent periods with distinct characteristics. Relying on the chronology of the data, we can situate tweets within an unfolding conversation which tells us dually about the events at hand as well as variations in tweeting practices.

Corpus analysis of semantically driven data

A **corpus** (or topic-based) view of **semantically driven** Twitter data aims to uncover the semantic makeup of a whole dataset. This is achieved in Chorus using the term co-occurrence visualisation model (Cluster Explorer). This mode of analysis affords the discovery of sub-topics and themes around an original topic set by keyword criteria.

In order to explore these topical clusters further, we make use of 'cluster maps' (Figures 2, 3, 4 and 6), which place terms occurring together frequently within tweets

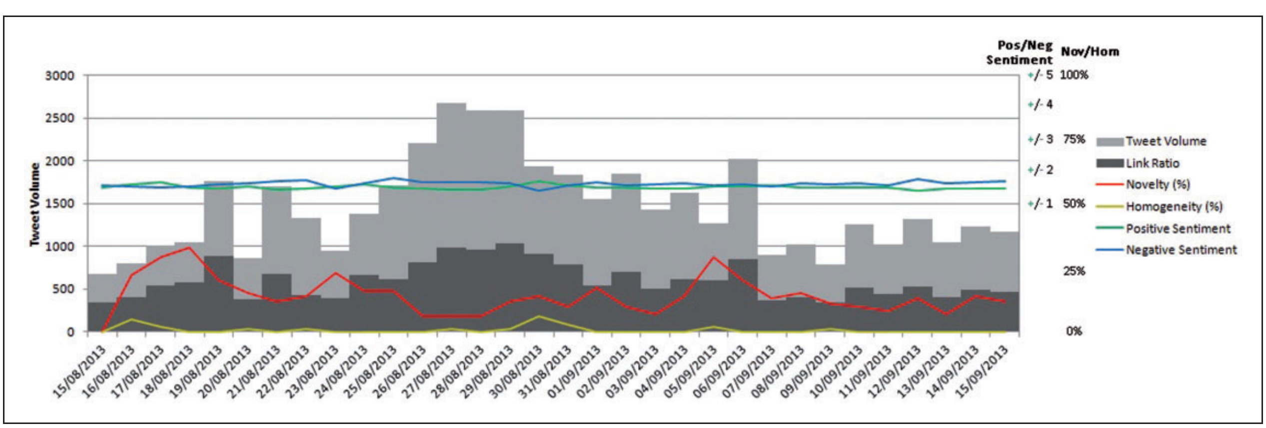

Figure 5. Timeline view of 15 key badger activists' timelines.

Downloaded from by guest on July 18, 2016

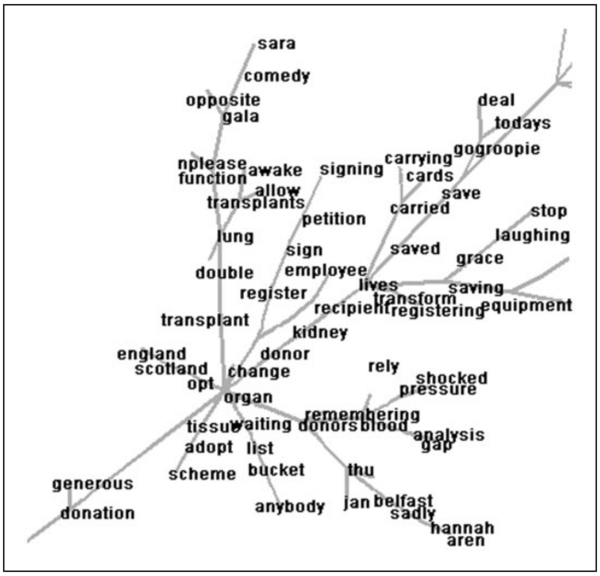

Figure 6. Clustering around the term 'organ', taken from Dataset 1.

closer together in space. Using this distance-similarity metaphor causes topics to emerge as structures (clusters, hubs, branches, etc.) within the information space of the map. Whereas a temporal view is event based (due to the choice to view data in discrete intervals) a corpus view is *topic based*, allowing researchers to delve deeper into the subjective content of tweets.

Example: User experiences of epinephrine auto-injectors ('epipens'). Our example for this analytic mode is around the topic of epipens (query keyword: 'epipen'): a popular brand of hand-held medical device for administering epinephrine in the event of an allergic reaction. Although the data collection method is the same (keyword searching) this dataset is most appropriately treated in a different way to the e coli example presented above. In this instance we were not aware, a priori, of any important chronological aspects of interest – rather, we wanted to explore everyday epipen user-experience issues. It is a relatively low-volume dataset (~4000 tweets over 68 days) with little convergence on a single sub-topic.

Zooming in on one section of the term-level map (see Figure 2) reveals a branch oriented around discussion of user experiences of epipen devices, diverging from a larger central node (top left of Figure 2).

Inspection of these branches indicates a discussion about accessories for carrying epipens. Tracing the discussion back to the 'root term' at which the topic diverges from other threads, we see the terms 'leg-buddy' and 'waistpal' (referring to products for making carrying an epipen convenient and subtle) are key to the formulation of a distinct 'devices and user-experience' branch (Figure 3).

Figure 4 shows tweeters associating terms pertaining to physical aspects of epipens with the term 'cases'. Here we see that the user experience of epipen cases is largely negative and relates to their size and visibility (and accordingly, the sizes of devices), e.g.:

> Not fucking pleased with this new epipen, its MASSIVE, needs an even bigger case...how am I meant to sit down???
> I am allergic to nuts carry an epipen AT ALL TIMES but am a certified swim coach and wear a one piece... my life sucks[8]

This mode of analysis allowed us to explore a broad topic of interest – epipens – without relying on simple term frequency to point us in any particular direction. Navigating around the cluster map in this way, analysts can sift their data for 'needles in haystacks' – here, this provided insight into user experiences with epipens unlikely to be uncovered with more formal search terms (i.e. 'weight' and 'size').

Temporal analysis of user-driven data

A **temporal** (or event-based) view of **user-driven** data captures a diverse array of interests in the selected user group, with analyses relying on the chronology of the data to elicit a narrative of how various features – e.g. term frequencies, usage of URL links, sentiment, novelty and homogeneity of conversation, etc. – fluctuate over time. The people whose timelines we capture may display interests in areas other than those we selected to be users to exemplify; these too become available to us. The data thus represents a proliferation of themes within which the topic of interest is embedded. The analytic focus is primarily in the periods where the user group converges on or diverges from some issue or event. In this way, a temporal view of user-driven data pieces together a story describing a set of evolving issues expressed by a user group and digs into that data beyond linguistically oriented accounts of frequencies of key terms.

Example: bTB and badger activism in the UK. The example we present here concerns a period of activist activity around UK proposals to cull badgers in countryside areas to prevent the spread of bTB. Our search strategy captured the Twitter timelines of 15 most frequent users of a selection of hashtags[9] through which badger culling activism was expressed, yielding 46,494 tweets from 15 August 2013 to 15 September 2013. This data reflects the entire Twitter output of these users over this period, whether tweets pertained to badger culling or otherwise. Our objective was to explore the practices through which activists mobilise Twitter in their activism.

Downloaded from by guest on July 18, 2016

The time period captured includes a key moment: the announcement on 27 August 2013 of Defra's badger culling programme. The most voluminous interval in the dataset (Figure 5) occurs on 27 August 2013 itself, during which the novelty metric is at its lowest – for the few days leading up to the Defra press release tweets became highly convergent around the cull announcement. Furthermore, not every tweet contains original content – often, the activists tweet ideas and links multiple times. This is reflected in the significant increase in the homogeneity metric on 30 August 2013 which saw a convergence on certain terms to propagate their message – in addition to obvious terms like 'badger' and 'cull', terms like 'make', 'save' and 'iTunes' were used to engage non-activists in the debate (i.e. with encouragement to 'save' badgers through signing an e-petition, and to purchase Brian May's new anti-cull charity single via 'iTunes' in an attempt to 'make' it chart).

Despite their concern with supporting badger activism, the 15 selected users do tweet about different topics. Visually, these are seen as the red novelty metric rises in the timeline, as well as 'crossovers' where the typically overriding negative (blue) sentiment dips and the typically lower positive (green) sentiment rises. In periods of spiking novelty, we find tweeters turning to new topics, e.g. reports of dolphin killing in Taiji (24 August 2013) or fracking (18 August 2013). Where sentiment 'crossover' points occur in conjunction with novelty spikes, these periods denote a change in the net positivity and negativity of words used to express newly introduced topics. We see this happen on 5 September 2013 when activists celebrate the birthday of Freddie Mercury – with Brian May as the celebrity figurehead of the movement, his ex-Queen bandmate Freddie Mercury's birthday is celebrated as a way of showing support for May.

Overall, our 15 activists express different interests that are not wholly disconnected from the badger culling debate, yet neither do they constitute a part of it; we begin to get a sense of their broader interests as 'activists' within multiple environmental issues. It is clear that badger culling activism dominates their talk, but using user-driven data we can do more to understand the broader personas of those people whose practices constitute this conversation.

Corpus analysis of user-driven data

A **corpus** approach to **user-driven** data collection sidesteps any lack of a priori knowledge as to how people tweet about a given topic on Twitter. There are research tasks for which effective query keyword criteria cannot be ascertained beforehand. Hence, the purpose of adopting this mode of analysis is to explore the overall topical makeup of the dataset to *find out* what kinds of things a user group tweet about, using a cluster map showing connected terms of interest.

Example: Symptom reporting of cystic fibrosis sufferers (and families of sufferers). Here, we used Chorus' data collection tool to capture user-driven data from a selection of followers of a cystic fibrosis (CF) news account. The number of followers at the time of collection exceeded 6000 and yielded a total of over 3,000,000 tweets over an approximately six-month period (14 February 2013 to 23 August 2013). To make the analysis more tractable, we filtered the dataset by selecting tweets from the lower end of the tweets-per-day spectrum (ranging from 0.01 to 29.36). Our analysis here focuses on the first 1797 users, who tweet between 0.01 and 0.61 times per day on average and together yield a total of 282,129 tweets. This was further broken down into two 'half' datasets of 141,063 (Dataset 1) and 141,066 (Dataset 2) to alleviate the computational load associated with processing the visualisations. Our interest here was in locating and understanding sufferers' reports of the everyday experiences of CF, to identify issues of importance which may go unreported in formal medical interactions.

This approach allowed us to discover topics of interest to our user group of candidate CF sufferers that fell outside of our expectations. Exploring the cluster map revealed a varied array of topics, reflecting the everyday nature of the users' conversations captured. However, noticeable clustering occurred around a key topic pertaining to the term 'organ' and the connected terms 'double', 'lung' and 'transplant'.

This clustering conveys a picture of transplant talk as having a significant relation to lungs – this much might be expected amongst candidate CF sufferers. Having identified this cluster we were then able to drill further down and found a distinct set of tweeting practices around the topic. Here, tweeters routinely involved themselves in personal communications expressing and receiving concern for CF sufferers known to be awaiting or undergoing double lung transplant surgery and recovery, e.g.:

> @ConcernedTweeter thankyou :) Yeah I'm needing a transplant badly now, still fighting everyday though! #organdonation #CysticFibrosis
> RT: will every cfer please keep @CFSufferer in ur prayers, she's in theatre right now getting a double lung transplant!
> @CFSufferer I hope you are doing unreal since the transplant! Was so delighted to hear the news #wooohoo

Downloaded from by guest on July 18, 2016

These same tweeters also utilised transplant surgery episodes to topicalise important related issues (such as post-operation aftercare and the organ donor register), e.g.:

> @CFCharity it's my sis's 30th bday today. She has CF & had a double lung transplant 1 yr ago which saved her – need more awareness!!!
> RT: @CFSufferer Its transplant week next week. I'm alive because of an Organ Donor. Please sign the organ donor register! #RT

Aside from their interpersonal communication, these tweeters make active use of the publicly visible nature of Twitter to help encourage others to recognise the emotive nature of transplants for CF sufferers and to campaign for positive action (i.e. registering as an organ donor). Our topic-based approach unveiled a cluster of key issues which would be difficult to locate with keywords, given the term 'transplant' is likely used more widely on Twitter than we would find relevant to candidate CF sufferers specifically. We were then able to investigate what this topic consists of for the selected user group and explore how the topic is structured and achieved through those users' tweeting practices.

Selecting a strategy

Given the different characteristics of the two modes of data capture outlined above, it is useful to review the reasons for choosing one over the other. Semantically driven data collection is suited to conversations where some unifying (set of) term(s) is known already and reflects people's usage of terms (rather than artificially *creating* a topic by filtering data with keywords). Given the focussed nature of this data, it is well placed to provide insight into broader trends – e.g. in predicting election results (Tumasjan et al., 2010). In contrast, user-driven data is more sensitive to the variety of different topics that specified groups of users tweet about. User-driven data is less focussed than its semantically driven counterpart, but enables researchers to inductively derive relevant keywords and topics. The decision about which strategy to adopt should be a data-driven process dependent on the research question. This requires experimentation with different data collection and analytic methods – having had a 'hands-on' approach to collecting data, researchers will find themselves equipped with better understandings of how to treat that data analytically.

Similarly, analytic work should start with a period of exploration to ascertain whether the data lend themselves to an event- or topic-based analysis. Initial visualisations and summaries of the data are revealing – are there distinct events, and what interesting things might be said about them? Or does the dataset show a corpus of topics for which chronological ordering does not produce insightful findings?

This process of exploration may keep iterating across any or all of the four cells outlined in Table 1, the end result being that researchers will find themselves with a set of research questions, a dataset which reasonably contains answers to those questions, and an analytic approach for drawing out those answers. This iterative process is the essence of visual analytics. We have demonstrated the value in applying visual analytics to social media research projects by positing four empirical examples as initial steps upon which deeper iterations might be built. An example of how we envisage this working: our corpus analytic user-driven study of cystic fibrosis sufferer experiences uncovered a keyword – 'pwcf', or 'Person/People With Cystic Fibrosis' – which we might feasibly go on to use as the basis for a query keyword search to see how topics around the term 'pwcf' change over time (i.e. a temporal view of semantically driven data). Unfortunately, for present purposes we have had to refrain from the iterative work of 'switching cells' in our examples, instead posing one example per cell so as to clearly demarcate each approach. Nonetheless, we hope readers will appreciate the value in iterating across the space of the framework.

Conclusion

We present a set of complementary methodologies for undertaking analyses of Twitter data as a socio-technical assemblage, with the emphasis on navigating around and unpicking the factors that construct and constrain the data. This notion of the research process as engaged in the production of assemblages informs this paper from top to bottom. To achieve this, we have taken a visual analytic approach (Thomas and Cook, 2005) wherein visualisations are utilised as tools for forming and pursuing hypotheses rather than results in themselves. Given our abductive grounding, this exploratory focus is highly appropriate, in that it is conducive to developing and defending interpretive accounts of social media in data-led ways.

Inasmuch as methods and methodologies are only as valuable as the empirical results they may yield, we can expect different social media projects to require new methodologies to support different modes of data collection and analysis. Given that our approach to building the visual analytic methodology is partly shaped by Chorus – a text-based Twitter analytics suite – as an element of our own research assemblage, the scope of our work is bound by the specific affordances Chorus provides. Thus, it is misleading for us to profess to have insight into how visual analytics might apply to projects

Downloaded from by guest on July 18, 2016

Chorus cannot currently support (i.e. on non-'microblog' platforms or with non-textual facets of Twitter data). However, we hope to have demonstrated that the general idea of using visualisations as tools for exploring data assemblages stands as a provocative alternative way for researchers to use existing tools to work with their data differently. We have focussed on the utility of visual analytics for text-based Twitter data in the hope that others may take up the reins and modify those principles to fit other platforms and data. We anticipate that our delineations of semantically and user-driven data and temporal and corpus analyses might be useful in this regard, as a demonstration of a framework for helping researchers think about and organise their research, and to create a foundation for further thinking around possible applications of visual analytics throughout digital social science generally.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Current estimates indicate that Twitter's output exceeds 340 million tweets per day from 332 million active users (http://en.wikipedia.org/wiki/Twitter, accessed on 06/05/2016).
2. Tufekci (2014) does express reservations about how far a singular focus on Twitter might take the field, in terms of such things as unrepresentativeness and skewing the direction of research. Nonetheless, Tufekci acknowledges the value in a paradigm which encourages a community research effort around shared datasets, tools and problems.
3. www.chorusanalytics.co.uk.
4. Specifically, Chorus' data collection routines draw on the following methods within Twitter's REST API. Query keyword searches use GET/search/tweets. User timeline retrievals use GET/statuses/user_timeline to provide tweets, with GET/friends/list and GET/followers/list methods to build lists of users to follow (though Chorus also allows for users to provide their own lists of tweeters to follow).
5. It is worth noting here that although they are associated, the metrics in these pairs – positive and negative sentiment, and novelty and homogeneity – do not necessarily negatively correlate as might be expected. Positive and negative sentiment utilise the SentiStrength algorithm (Thelwall et al., 2010), ascribing sentiment values to terms within tweets, and tweets may feasibly contain strong positive and negative terms simultaneously (e.g., 'I love tea but hate coffee'). Similarly, novelty and homogeneity are not necessarily inversely related – novelty detects shifts in word usage *between* an interval and intervals immediately preceding it, whereas homogeneity reflects the extent to which tweets *within* an interval tend to use the same terms. In this way, an interval may show both a high novelty and homogeneity value, i.e. tweeters may be using a relatively small vocabulary within an interval (high homogeneity), though their talk in that interval may be markedly different than the talk in previous intervals (high novelty).
6. For brevity, our analysis terminates at interval 31.
7. We recognise that this may be an artefact of the data, in that a relatively low volume of tweets amplifies differences between intervals in terms of novelty/homogeneity – we make the point for demonstrative purposes.
8. Usernames have been anonymised and tweet content paraphrased to protect the anonymity of tweeters.
9. The hashtags and the usernames derived from them are not reproduced here since this could compromise these users' anonymity.

References

Anderson CW (2011) Deliberative, agonistic, and algorithmic audiences: Journalism's vision of its public in an age of audience. *International Journal of Communication* 5: 529–547.

Blaikie N (2000) *Designing Social Research: The Logic of Anticipation*. Cambridge: Polity Press.

boyd d (2010) Social network sites as networked publics: Affordances, dynamics, and implications. In: Pappachrissi Z (ed.) *A Networked Self: Identity, Community, and Culture on Social Network Sites*. Abingdon, UK: Routledge, pp. 39–58.

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

boyd d, Golder S and Lotan G (2010) Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In: *Proceedings of the 43rd Hawaii International Conference on System Sciences*, Hawaii, USA, 5–8 January 2010, pp.1–10. IEEE Computer Society.

Brügger N and Finnemann NO (2013) The web and digital humanities: Theoretical and methodological concerns. *Journal of Broadcasting and Electronic Media* 57(1): 66–80.

Bruns A and Burgess JE (2011) The use of Twitter hashtags in the formation of ad hoc publics. In: *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference*, Reykjavik, Iceland, 24–27 August 2011.

Burnap P, Williams ML, Sloan L, et al. (2014) Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining* 4(1): 1–14.

Card S, Mackinlay J and Shneiderman B (1999) *Readings in Information Visualization: Using Vision to Think*. New York: Morgan Kaufmann.

Cui W, Lin S, Tan L, et al. (2011) TextFlow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics* 17(2): 2412–2421.

Downloaded from by guest on July 18, 2016

Diakopoulos N, Naaman M and Kivran-Swaine F Diamonds in the rough: Social media visual analytics for journalistic enquiry. In: *IEEE Symposium on Visual Analytics Science and Technology*, Salt Lake City, UT, USA, 25–26 October 2010, pp.115–122. IEEE.

Gillespie T (2014) The relevance of algorithms. In: Gillespie T, Boczkowski PJ and Foot KA (eds) *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge, MA: MIT Press, pp. 167–194.

Hassan S, Sanger J and Pernul G (2014) SoDA: Dynamic visual analytics of big social data. In: *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, Bangkok, Thailand, 15–17 January 2014, pp.183–188. IEEE.

Heverin T and Zach L (2011) Use of microblogging for collective sense-making during violent crises: A study of three campus shootings. *Journal of the American Society for Information Science and Technology* 63(1): 34–47.

Hughes J and Sharrock W (1997) *The Philosophy of Social Research*. Harlow, UK: Pearson Longman.

Kitchin R (2014) Big data, new epistemologies and paradigm shifts. *Big Data and Society* 1(1): 1–12.

Langlois G (2011) Meaning, semiotechnologies and participatory media. *Culture Machine* 12: 1–27.

Locke K (2010) Abduction. In: Mills AJ, Eurepos J and Wiebe E (eds) *Encyclopedia of Case Study Research*. London: Sage Publications, pp. 1–3.

Luo D, Yang J, Krstajic M, et al. (2012) EventRiver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics* 18(1): 93–105.

Mahrt M and Scharkow M (2013) The value of big data in digital media research. *Journal of Broadcasting and Electronic Media* 57(1): 20–33.

Marcus A, Bernstein MS, Badar O, et al. (2011) TwitInfo: Aggregating and visualizing microblogs for event exploration. In: *CHI '11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada, 7–12 May 2011, pp.227–236. New York, USA: ACM.

Marshall C (2012) Big Data, the crowd and me. *Information Services and Use* 32(3–4): 215–226.

Matthews N and Sunderland N (2013) Digital life-story narratives as data for policy makers and practitioners: Thinking through methodologies for large-scale multimedia qualitative datasets. *Journal of Broadcasting and Electronic Media* 57(1): 97–114.

Murthy D (2012) Towards a sociological understanding of social media: Theorizing Twitter. *Sociology* 46: 1059–1073.

Murthy D (2013) *Twitter: Social Communication in the Twitter Age*. Cambridge: Polity Press.

Raghavan P (2014) It's time to scale the science in the social sciences. *Big Data and Society* 1(1): 1–4.

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

Rose S, Butner S, Cowley W, et al. (2009) Describing story evolution from dynamic information streams. In: *IEEE symposium on visual analytics science and technology*, Atlantic City, NJ, USA, 12–13 October 2009, pp.99–106. IEEE.

Sharma S (2013) Black Twitter? Racial hashtags, networks and contagion. *New Formations* 78: 46–64.

Thelwall M, Buckley K, Paltoglou G, et al. (2010) Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12): 2544–2558.

Thomas J and Cook K (2005) *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Richland, WA, USA: Pacific Northwest National Laboratory.

Tufekci Z (2014) Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, Ann Arbor, Michigan, USA, 1–4 June 2014, pp.505–514. Palo Alto, CA: The AAAI Press.

Tumasjan A, Sprenger TO, Sandner PG, et al. (2010) Predicting elections with Twitter: What 140 characters reveals about political sentiment. In: *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Washington, D.C., USA, 23–26 May 2010, pp.178–185. Menlo Park, CA: The AAAI Press.

Downloaded from by guest on July 18, 2016