Article

new media & society
2018, Vol. 20(9) 3341–3358
© The Author(s) 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1461444817748953
journals.sagepub.com/home/nms
SAGE

Understanding the production and circulation of social media data: Towards methodological principles and praxis

Susan Halford, Mark Weal, Ramine Tinati, Les Carr and Catherine Pope
University of Southampton, UK

Abstract

Social media data have provoked a mixed response from researchers. While there is great enthusiasm for this new source of social data – Twitter data in particular – concerns are also expressed about their biases and unknown provenance and, consequently, their credibility for social research. This article seeks a middle path, arguing that we must develop better understanding of the construction and circulation of social media data to evaluate their appropriate uses and the claims that might be made from them. Building on sociotechnical approaches, we propose a high-level abstraction of the 'pipeline' through which social media data are constructed and circulated. In turn, we explore how this shapes the populations and samples that are present in social media data and the methods that generate data about them. We conclude with some broad principles for supporting methodologically informed social media research in the future.

Keywords

Data, method, methodology, population, sample, social media

Introduction

The phenomenal growth of online social media is a significant feature of the past decade. Social media are now firmly embedded across economies, cultures and political

Corresponding author:
Susan Halford, Web Science Institute, University of Southampton, Highfield, Southampton SO17 1BJ, UK.
Email: susan.halford@soton.ac.uk

processes and in the lives of hundreds of millions of people, most famously a billion users signing on to Facebook in 1 day. It is not just the new forms of social practice associated with social media that are extraordinary – or their consequences, shaping social relationships, politics and business models, for instance – but that these activities create continuous digital traces, routinely captured as data: a remarkable new source of social information.

Twitter data, in particular, have attracted widespread interest from business and government, but the response from researchers has been mixed. On one hand, large volumes of data can be 'harvested' easily and freely from the platform Application Programming Interface, offering evidence of the things that people say and do, in real time, over time and at scale. Not least at a time when research funding is constrained (certainly in the United States and United Kingdom), these data appear as gift, with rich potential to sustain and even extend social research. Indeed, recent years have seen significant take-up of Twitter data for social research across the computational sciences as well as the social sciences. On the other hand, there are concerns that these data are problematic, flawed by demographic biases and unknown provenance. Social research has been built on well-established principles grounded in clearly understood populations, carefully controlled sampling and well-known methods for data collection. Twitter data offer none of these. Accordingly, there are fears that the use of Twitter and indeed other social media data may lead to poor research and unsustainable claims, damaging the reputation of social research (Goldthorpe, 2016; Hardaker, 2016).

This article seeks to trace a middle path in the space between '*giving in and getting out*' (Gehl, 2015: 148). Our way forward is methodological. Working with conventional sources of data, professional standards demand that we make the details of our research design, methods of data collection and data management explicit and outline any relevant challenges in this regard. These discussions are rare in social research with Twitter data, where little consideration is given to methodological questions about the data, as we show in section 'Methodological implications'. Harvesting data is one thing, but understanding these data is quite another. To be fair, this is difficult to do. Twitter, like most popular social media platforms, is privately owned and makes data available on its own terms and with varying degrees of methodological information. Furthermore, the processes of data generation are technically complex and subject to frequent changes, which makes it challenging to track data provenance. However, just because Twitter offers a novel, opaque and dynamic source of secondary data, it is no less reason to consider methodological questions. To the contrary, there is all the more reason, if these, and other social media data, are to be a credible and sustainable source for research.

This article offers a practical approach, by beginning to unpack the methodological challenges of social media research and considering how best to respond to these. In section 'What are social media data?', we ground our approach in Science and Technology Studies, long used to conceptualise data infrastructures (Bowker, 2005; Bowker and Starr, 1999) and influential in recent theorisations of the broader 'dispositifs' (Ruppert et al., 2013) or 'assemblages' (Kitchin and Lauriault, 2013) that produce new forms of digital data. This demands that we understand data as the outcome of the activities of heterogeneous actors, from databases, interfaces and browsers to consumers, markets and legal regulations. However, while this provides a rich and promising conceptual framework to underpin social media research, it offers little in the way of practical methodological guidance. To redress this, we propose a sociotechnical conceptualisation of the 'data pipeline' that shapes the construction and circulation of Twitter and other social media data. In section 'Methodological implications', we consider the methodological consequences of this pipeline, examining how it shapes the population, sample and methods of Twitter and other social media data construction and circulation. To do this, we draw on technical experiments conducted with Twitter data sets. While these have been reported as research findings in themselves, they are rarely taken into account in social research that makes use of Twitter data, as we show throughout section 'Methodological implications'. While Twitter is our main focus in this article, our pipeline conceptualisation and the methodological questions that we raise also apply to other social media platforms, and we illustrate this throughout.

Overall, our aim in this article is to begin a systematic discussion about how Twitter and other social media data are shaped for secondary research analysis. In doing so, we offer some practical guidance based on experiments that we hope will support the development of a new methodological praxis for Twitter and other social media data. We are driven by a core philosophical research principle: we should understand and describe our data – what we do *and do not* know about them. This is not to insist on a gold standard for research data, whereby we must have full knowledge of data provenance, or to offer a finite description of data construction and circulation for any particular social media service. Indeed, neither is possible partly because we cannot access all the information that would be necessary and partly because social media services – and the data available from them – are highly dynamic. But it is to insist on methodological rigour. Section 'Discussion and conclusions' suggests some key methodological principles for the use of social media data that might strengthen – and thereby protect – this new source of data for social research. Our conviction is that this will produce better academic research and will also develop our critical capacity to contribute to, and where necessary critique, the claims that are increasingly made from social media data by governments, the media and other commercial organisations.

What are social media data?

As we enter the era of Big Data, Bowker's (2005) now emblematic statement that '*"raw data" is ... an oxymoron*' (p. 184) was never more apposite. As routine activities – from travel and shopping to web browsing and, of course, social communication – generate data of unprecedented volume, variety and velocity some dramatic claims have been made that these data constitute 'the new oil' (Humby, 2014), a new natural resource that will, at last, reveal the mysteries of the social world (Anderson, 2008; Mayer-Schönberger and Cukier, 2013; Watts, 2011). Conversely, our starting point is that these data are, like all data, '*always already social*' (Bowker, 2013: 168). Data do not exist 'in the wild' but are constructed rather than discovered (Manovich, 2001) through a network of activities involving both human and non-human actors: social scientists and users, concepts and categories, survey tools, statistical measures, publication infrastructures and so on (Hacking, 2006). As Gitelman and Jackson (2013) argue, data are both framed – actively constructed in specific contexts – and framing – themselves constructing objects and subjects of knowledge.

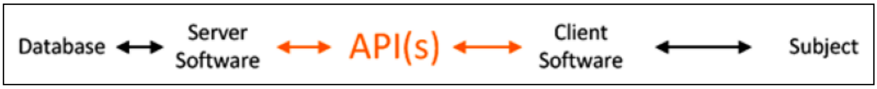

Figure 1. Basic pipeline of social media data construction and circulation.

These conceptual observations apply to *any* data but acquire particular significance in the current context, where researchers are drawing on new forms of digital data, not of our own making. In doing so, we make ourselves '*... reliant on platforms, methods, devices for data processing that have been developed in contexts and for purposes that are in many ways alien to those of social research*' (Marres and Weltevrade, 2013: 13).

This is not social research 'as usual'. These data are generated beyond the orbit and control of researchers, used to producing their own data and/or working with carefully described secondary sources of data. In contrast, social media data are constructed and owned by commercial companies, for whom the data are their only asset (Burgess and Bruns, 2012): an asset that is carefully protected and shared under conditions of the companies' choosing and within relevant legal frameworks. Reflecting on this, Ruppert et al (2013) call for attention to

> … the lives and specificities of devices and data themselves, where and how they happen, who and what they are attached to and the relations they forge, how they get assembled, where they travel, their multiple arrangements and mobilizations and, of course their instabilities, durabilities and how they sometimes get disaggregated too. (pp. 31–32)

The construction and circulation of social media data involve a heterogeneous network of actors. As with any network, there are many places that we could begin. Since we are concerned with methodological questions, our proposal is that we start with the 'pipeline' of data construction and consumption. Conventionally used in Computer Science, this metaphor describes the linear processes that shape the technical management of data (Patterson and Hennesey, 1998). In what follows, we corrupt this metaphor in two ways. First, we understand the processes shaping the construction and circulation of data along the pipeline to be social, political and economic, as well as technical. Second, we understand these processes as relational and dynamic, rather than fixed or necessarily linear.

Figure 1 provides an abstraction of actors in the 'pipeline' of social media data construction and circulation: the subject who creates the content, posting to a social media platform, through client software on a phone, laptop and so on that represents the data to the Application Programming Interface(s) (API), which enforces rules to determine what is passed through to the company's server software, and how, and the server software that organises content into databases that store data in particular formats and structures. In turn, this shapes if and how these data are shared with users – including researchers – back down the pipeline. Furthermore, the 'output' is *not* a simple reversal of the 'input', but is shaped by the methods that researchers use to access data (which may introduce new actors into the pipeline, including data re-sellers; see Figure 2), the economics and practicalities for the companies in sharing data, with whom and on what basis, both shaped by legal and ethical considerations.

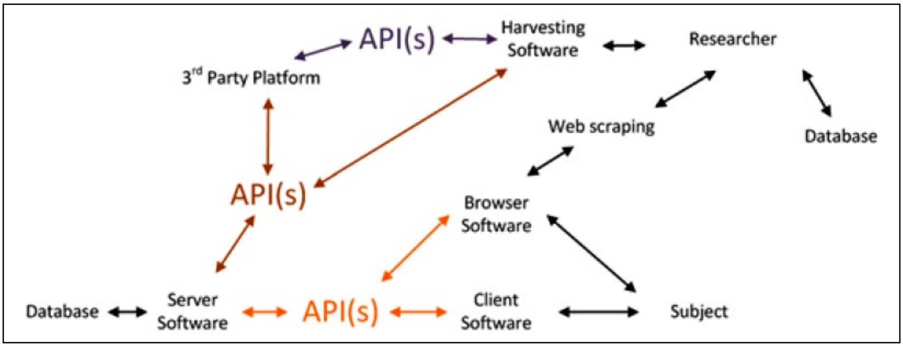

Figure 2. The research data pipeline.

Methodological implications

The processes described in the data pipeline have profound methodological implications for use of Twitter and other social media data. However, to date, they have rarely been considered in research that uses these data for social research. In May 2016, we conducted a short survey of social research using Twitter data and published in leading journals in Social Sciences and Computer Science between 1 January 2013 and 31 December 2016. Our search strategies were adapted to the different disciplinary fields, using (1) the International Bibliography of Social Sciences and (2) contents of the three highest ranking (H5) social network journals in Computer Science[1] (which has no unified database of publications). We searched for ‘Twitter’ in the title or abstract, seeking papers where Twitter was a substantial element, rather than a passing reference. IBSS returned 1849 papers and the Computer Science journals returned 58 papers. To construct a comparable sized sample from the Social Science journals, we selected the 57 papers from the highest ranking Social Science journals on our list[2] (also ranked by H5). Of the 115 papers, 90 had used Twitter data. These were hand-coded for information about (1) the method of data collection, (2) the population – who or what was in the data and what steps if any were taken to clean data, and (3) the shaping of data by platform functionalities.

Our findings demonstrate that little attention is paid to the pipeline of Twitter data construction and circulation in this social research. Nonetheless, as we conceptualise above, the implications may be significant. In what follows, we will examine this more closely by focusing on how the pipeline shapes Twitter and other social media data along three key methodological dimensions: the population, the sample and the method of data construction. In each case, we return briefly to the findings of our survey.

Population

In broad terms, all social research begins by scoping of the ‘population’ to be researched: defining the empirical subject. Most commonly, social researchers think of this in terms of people, their characteristics, values and actions. In this respect, the appeal of social media data is that they offer insights into the everyday lives of their ‘users’, the subjects who post content online, situated at the far right of the pipeline. We know already that social media users are skewed sub-sets of a global population. In a world where over half the population does not have access to the Internet,[3] it could not be otherwise. Even among the 3.7 billion Internet users worldwide, while the number of social media users is impressive, it is no basis for making claims about the total population. Surveys tell us, for example, that Twitter users are middle class,[4] that more women than men use Pinterest[5] and that 70% of WhatsApp users are aged under 45.[6] These surveys are important because not all social media platforms reveal demographic information to researchers (Facebook does not) or even collect demographic details (anonymity is a key feature of Yik Yak; WhatsApp accounts are defined by a telephone number, not an individual). Where demographic information is available, it is not necessarily evident if or how this is related to offline characteristics. Social media companies take different positions on this: Facebook famously barring those it deems 'inauthentic' and Instagram and Twitter taking little interest in this.

Location is also a case in point here. Mapping social media data is hugely popular (Doré et al., 2015; Leetaru et al., 2013; Rodríguez-Amat and Brantne, 2016), accelerated by mainstreaming of location-based functionalities from specialist platforms (e.g. Foursquare) into the big social network platforms (Evans and Saker, 2017). However, there has been little attention to how geolocated data are constructed along the pipeline. Users may add their location manually (e.g. to their profiles), which may be more or less accurate. For example, there is evidence that the Iranian diaspora on Twitter selected Tehran as their location during the Iranian elections, in a show of solidarity, while political activists in Iran chose to hide their location at the time (Gaffney, 2010). Alternatively, users may enable their client software to add location to the metadata attached to each post. This is likely to be a more accurate method, but experiments suggest that fewer than 2% of Twitter users select this option and that fewer than 3%[7] of tweets contain geotagged metadata (Leetaru et al., 2013). Furthermore, the volume of geotagged tweets has its own geography, with 2.86% in Jakarta compared to 0.77% in Moscow, and taking into account that some users are particularly active contributors to the overall Twitter stream, Leetaru et al. (2013) conclude that just '*… one percent of all users accounted for 66 percent of georeferenced tweets*' (n.p). We cannot assume that this small number of geotagged accounts is a representative of the wider Twitter population. The decision to disclose – or not – is likely to shaped by a variety of individual and social factors, biasing the geotagged population in significant ways.

Finally, much social media research assumes that each account belongs to an individual human actor. However, on some platforms (e.g. Instagram, Tumblr, Twitter and WhatsApp), there is no limit to the number of accounts that anyone can register, and data may include corporate, group and parody accounts. Corporate accounts are distinctively crafted to represent particular organisational interests, rather than personal views, activities and identities. Parody accounts (Highfield, 2016), for example, the Dark Lord (@LordVoldemort) and Elizabeth Windsor (@Queen_UK) on Twitter post thousands of tweets to their respective followers (currently 2.04 and 1.3 million), with retweets reverberating across the Twitter network. While some may be easily spotted from their qualitative features – not necessarily examined in quantitative analytics – others may be less obvious. Related to this, automated accounts make up an increasing proportion of activity and content on social media. These computationally controlled 'bots' may be new accounts created solely for the purpose or 'hijacked' accounts providing cover for bot activity. Either can be readily bought to add friends and followers, push posts across social media networks (e.g. liking or reposting) or aggregate content. Estimates of bot activity vary, from 4.6% on Sina Weibo (Zhang et al., 2016) to an official Twitter estimate of 8.5% and an unofficial estimate of 13.5% (Chu et al., 2010). In addition, it is estimated that as many as 38% of all Twitter accounts may be 'cyborgs' enhanced by computational agents (for example) pushing out queued posts at regular intervals or optimum times of the day (Chu et al., 2010). Bots may or may not be identifiable depending, in part, on their behaviours along the data pipeline. While the simplest way for bots to operate is to engage directly with the API (feeding activity in directly, rather than through client software), these may be relatively easily detectable and so deletable. Bots that simulate interaction with the user interface are less readily identifiable. Meanwhile, social media companies themselves may also use bots to drive up activity and enhance use of the platform.

In short, we cannot and should not assume that the social media population is representative of the wider population. More than this, the constitution of the social media population along the data pipeline shapes the nature of the data generated. Our survey of research using Twitter data shows that such issues are rarely considered. Only 9% of papers discussed the demographics of their data sample, 4% mentioned corporate accounts and 3% mentioned the presence of automated accounts. Data cleaning was mentioned in 6% of papers, but only one of these was concerned with user type (removing corporate accounts). While 25% were interested in geographically defined samples, there was little consideration of the mechanisms generating location data, their reliability or social selectivity.

Sample

Despite the often-made claim that big data provide total populations, ending our reliance on samples, this is rarely the case for social media data (Highfield et al., 2013). While, in principle, all activity is captured, whole data sets are rarely shared with researchers. Some are given, or buy, full data sets, but the vast majority are dependent on smaller samples, drawn from either their own web scraping or the use of a public API (either directly or indirectly through the services of a third-party data broker). Figure 2 elaborates the basic pipeline representation, with particular reference to the 'output' back down the data pipeline. The mode of circulation shapes the nature of derived data in ways that may be significant.

For example, web scraping uses automated agents (computational programs) to process web pages and extract specific pieces of information, for example, the content of social media posts. This is advantageous where social media services provide no formal access to their data through an API or if the researcher is seeking data other than those officially provided (although this may contravene company terms and conditions). Web-scraped data have some distinctive characteristics. Most notably, the agents are requesting pages created by other agents – Google, for instance – whose algorithmic processes may vary the results according to the characteristics of the account, browser, computer, IP address and location of the request (the ‘filter bubble’ (Pariser, 2012)). Any content received through web scraping is, thus, already sampled. Furthermore, web-scraped data will only include whatever information is available to the browser, which may be different to data sourced directly from the companies. Although additional information may be inferred through web scraping – for instance, linking to other data on the web pages searched – this will require particular assumptions and inferences to be built into the search processes.

Alternatively, data may be harvested directly from the social media companies. While full data sets may be commercially available, sometimes public APIs offer sampled data for free. How this sample is structured may have significant methodological implications. Let us take Twitter as an example. Launched in 2006, Twitter was initially open about sharing data, but as the company moves ‘*... from many possibilities to a narrower commercial entity*’ (Burgess and Bruns, 2015: 97), access has been progressively restricted and third-party data brokers (adding their own functionalities to data streams) have largely been subsumed within the company (GNIP was bought by Twitter in 2014) or had their data access restricted (DataSift lost access to the full live Twitter data stream and to historical data in 2015). While Twitter pursues a model of commercialisation for its most valuable data streams, access to other data streams is still available through the public API or, rather, two different APIs: the streaming (live) API and Search API, with ‘*[e]ach offer[ing] a different set of methods for interacting with the system and each constrains the user in different ways*’ (Driscoll and Walker, 2014: 1748).

At present, the Streaming API provides real-time data in two ways: (1) a 1% sample of all tweets, ‘pushed’ through the API on a continuous basis. We do not know how this sample is generated, but the company states it is random (perhaps a time-stamped sample[8]), and this is validated in experimental work (Morestatter et al., 2013; Wang et al., 2015). This sample may be very useful for looking at ‘what is happening on Twitter’ but less so if the research aims to harvest data on a particular topic (most of which is unlikely to fall in to the 1% sample). (2) Here, the Streaming (*filter*) API allows users to harvest real-time data for specific search terms. This is likely to return a far greater proportion of the tweets for a given term (Gaffney and Puschmann, 2015), but there is no guarantee that it will return all tweets for that search term, even if these constitute less than 1% of the Firehose. Twitter also offers a Search API ‘*... to search against a sample of Tweets published in the previous 7 days*’ (https://dev.twitter.com/rest/public/search) (until recently this was 14 days). The sample received is ‘*focussed on relevance not completeness*’ (Gaffney and Puschmann, 2015), but we do not know how this is sampled. Experiments suggest that the Search API returns far fewer tweets than Streaming (filter) API, at a ratio of approximately 1:4 (Gonzalez-Bailon et al., 2012), while our own experiments confirmed this and showed far fewer retweets in the historic data sets.[9] Furthermore, Driscoll and Walker’s (2014) suggest that Search API data are skewed heavily towards central users and more clustered regions of the network. Meanwhile, the amount of data received may also be shaped by the client software (see discussion at https://news.ycombinator.com/item?id=4795052, accessed 5 October 2016).

In addition, Twitter and other social media platforms impose rate limits on the number of calls that can be made to an API during a given timeframe. This may be driven by practical limitations, for example, to load manage the network, it could be a business decision to stratify the service offered (pay more get more), or it might be the result of decisions – ethical or otherwise – about what to make public. Only 180 queries can be made to the Search API from each Twitter account in every 15 minutes (as of August 2016). On Facebook, each user can make 200 calls per hour, while the Instagram API (available at a cost) allows 500 calls per access token in a 1-hour timeslot. Whether this matters depends on the nature of the data being queried. Small, regular data streams may not be affected at all by rate limits, whereas large data streams will be incomplete and clustered in the first part of time windows. In either case, it will not be clear what percentage of the potential data available has been returned, unless there are also robust data supplied on the total number of messages sent during a given timeframe. Note, too, that web scraping methods are subject to a different form of rate limiting as there are restrictions on how many times a HTTP GET request can be called (the technical method used to retrieve Web data) before the server denies the request. More generally, the capacity of the social media company's servers at any given point in time may impact the amount of data delivered, while the geographical location of servers may affect the nature of data (if, for example, Safe Harbour arrangements do not exist between countries, it may not be possible to deliver personal data on individuals.

These descriptions are not intended as a technical 'manual'. Rather, they illustrate what the Twitter API *does* (Busher, 2013) to shape the data that are harvested by researchers. It is important to note that APIs evolve as new functionalities are added to the platform, as additions are made to underlying data models and as company decisions are made to change access, for example, with regard to rate limits. This has particular implications for research that seeks to replicate previous studies or to take a longitudinal perspective. However, returning to our survey, we found that 23% of papers offered *no description at all* of how the data were harvested, 46% stated that they used 'the Twitter API', but only 43% of these explained which API, and of these, few considered the implications of this for their data or findings. A further 24% of papers used data from third parties, including publicly available data sets and data broking services, and 45% used web crawling methods, but none considered the implications of these methods for their sample or research findings.[10]

Method

Social research has a rich repertoire of methods through which to 'capture' data, for example, questionnaires and interview schedules. Similarly, social media platforms are designed artefacts that record particular types of information and not others. Social media platforms present the world according to their designed features – posts, comments, 'friends' or 'likes' – and the emergence of associated cultural practices and norms of sociality. These are neither unconnected to the social, nor do they simply reflect an independent sociality. Social media functionalities show significant convergence across platforms over time, for example, the major platforms operate with a version of profile, timeline, followers/friends, likes/favourites and location. Using social data, we come to know the social through these features, retrofitting meaning to functionality, although, of course, what particular actions mean is far from evident.

For example, 'like' is a common feature, but motivations for and meanings of the action are not conveyed by the vocabulary of the interface. Indeed, Meier et al. (2014)

identify 25 different uses, ranging from indication that an item is topically relevant, acknowledging a family member, bookmarking, agreement with a statement, accident or trying to engage others. Furthermore, how the 'like' is added to the database may be significant, for example, whether we click a 'widget' on an online news source or a shopping website ('like us on Facebook') or whether this is a like in the user's own client device application. In sociological terms, these may indicate different things, but the data generated rarely distinguish between them.

Similarly, the number of account 'followers' may be used to indicate popularity and/or influence, but the meaning of 'following' is complicated. For a start, the more followers an account has, the more likely it is that these are bots, since bots are often programmed to follow accounts with greatest popularity and/or influence (Chu et al., 2012). Bots can be passive followers, significant if the focus is social influence, or more actively push information across networks, important if the focus is information diffusion. Meanwhile, even human followers are not guaranteed to actually read content posted to timelines, despite social media companies' investment in metrics to encourage users with (rather vague) information about 'impressions' or 'engagements'.[11] Moreover, not all user activity or information flow is captured by formal metrics, for example, users search and follow hashtags and keywords, and content gets shared through alternative channels.

Furthermore, functionalities, and our use of them, change over time. As users, we adjust our practices to social media platforms – doing things we may never have done before *and* over time, new practices may emerge, only possible because the platform is there, but that were never envisaged by the designers. Much Twitter research has focused on the 'retweet' function – to explore information flows and network formations. But prior to 2009, there was no formal retweet functionality, and until 2012, the Streaming API did not deliver retweets made with the retweet button because they were not identifiable in Twitter's internal data structure (Bruns and Stieglitz, 2012). At the same time, if the focus is information flows, there are plenty of other ways to pass on information. Indeed, boyd et al. (2010) suggest that 'dark retweets' (re-posts made not using the formal retweet convention or retweet button) may account for up to 40% of total re-posts and that these are domain specific, so our knowledge of information flows in particular parts of the network may be especially limited. The underlying data model and counting mechanisms may treat these actions as different, but whether this is so, and the nature of their significance, is far from clear.

Finally, the methods and design decisions of data management, through the organisation and configuration of servers and databases, may have significant effect on what data are returned to queries, shaping the types of analysis that can be performed. While data may be received as unstructured streams of user-generated content, engineers decide how they are stored and managed, with consequences for how they can be searched and are delivered in response to queries. Since Twitter and Facebook first launched their API, the richness and structure of the data made available have changed considerably. Twitter's data structure did not originally contain geolocation, retweet or hashtag data and has only recently incorporated the 'like' feature. Moreover, fields which have been present since Twitter first launched have also changed; the 'created_at' and 'text' field has gone through several iterations, with changes in format and markup. These changing schemas and data formats make it very difficult to assure consistency in data derived at different points in time, important if the aim is to replicate experiments or conduct longitudinal analysis.

Overall, Twitter and other social media data are constructed through specific methods and metrics of data collection and circulation. As Marres and Weltevrade (2013) have argued in a broader context, new forms of digital data '... *tend to come with external analytics already built in*' (p. 313), which requires reflection if we are to make the most of these data (see also Heer and Verdegem, 2015; Marres and Gerlitz, 2015). This was not considered in the papers that we reviewed.

Discussion and conclusion

This article has begun to develop a systematic account of the methodological challenges that arise when working with Twitter and other social media data. By conceptualising the 'data pipeline', we have drawn attention to the sociotechnical processes shaping the curation and circulation of social media data. We have made a particular point of explaining that the operation of technical actors is rarely considered by social researchers, and the experimental evidence of their effects on data has, to date, not been integrated into mainstream work with social media data. Overall, our contribution has been to extend the theoretical approach to data infrastructures (and associated accounts of assemblages, dispositifs, etc.) into a practical approach to social media data methodology with the aim of encouraging the emergence of new praxis that can establish these data on a firmer footing in the methodological cannon of social research. For ease of explanation, we presented the methodological challenges in a linear way, as they arise along the pipeline, beginning with the subjects on the far right and ending with the databases at the far left of Figure 1. So it is important to emphasise here that the processes along the pipeline are iterative: for example, changes in the client device may impact what users (can) do, changes in storage may impact how the API can be searched and – at the heart of data construction and circulation – changes to the API may impact along the pipeline in both directions, perhaps with rebounding effects, for example, as researchers turn to web scraping methods or new kinds of widgets are developed. Relatedly, we should mention that ethical challenges arise iteratively along the pipeline. For example, how should we treat personal data that users post on public pages? What data should social media companies release? What implications do data structure and format have for personal data linkage? and so on. In the papers we reviewed that included data, only three reported on completing an ethical review process and one other explained why ethical review had not been necessary. Taken together, this problematization of social media data may appear only to underscore the concerns expressed by those who have doubted their promise for robust social scientific research. This is not our intention. To the contrary, our tactic is to suggest that those of us using social media data should seek to address these challenges in our research. Certainly, we must accept that social media data are not like earlier generations of data, and consequently that the exact same methodological frameworks will not be appropriate. However, we should seek to position this new form of data methodologically and develop new frameworks that will ensure its future value for researchers. In this article, we proposed the 'data pipeline' approach to this, drawing our attention to how the construction and circulation of data shape population, sample and the method of data collection. This is summarised in Table 1.

Taken together, our survey and our own experience in the field of social media research suggest that the issues raised in this article are rarely considered. While we are critical of this, we are also sympathetic. Social media are a new source of data and none of us were, from the beginning, an expert. To use a Norwegian expression, we are all 'paving the road as we walk'.[12] Certainly, we recognise the omissions in our own work, as well as in the work of others.

Looking forward, we suggest three key steps towards a new methodological praxis, based on familiar principles. First is transparency, basic diligence in reporting how data are harvested, when, using which data streams and what search terms. In addition, as we would with any other method, we should record key metrics of the resultant data streams, including size and any other notable characteristics. These details matter in a number of ways. Most obviously, they underpin comparative and longitudinal research and the possibility of reproducibility. If the intention is to pursue such research, then we need to know whether we are comparing like with like, and if not what the differences are as well as if and how they might be significant. The API changes are a particular issue here, since these may have a significant effect on the data that are returned even to identical queries. While we will not always know what these changes have been, we will need to investigate these possibilities and/or – perhaps more likely – caveat the claims that we make accordingly. For example, if claims are being made with regard to a particular temporal phenomenon occurring, for example, the speed at which a comment is spread across the network, we should bear in mind where temporal representations have changed. This might include being aware that if the timestamp format and time zone record have changed in the Twitter data structure since the API was first made public, this means that comparisons between data sets harvested at different points in time may yield inaccurate and misleading results.

Second, it is important that we consider if and how data construction might matter for the particular research questions under consideration. Some of the issues that we have highlighted above will matter a great deal for some research questions, and not at all for others. For example, the presence of bots may not matter at all if the aim is to explore information diffusion across a social media network but may matter a great deal if claims are made about human forms of influence in these networks. Claims about temporal patterns of social media activity should bear in mind the potential presence of cyborg accounts, while geographical questions and mapping methodologies will need to consider the very low proportion of geotagged data and the potential biases of those who enable this. These considerations in turn may moderate the kinds of questions that are asked and the claims that can be made. Take the vexed issue of demographics, for example. One response to the well-known biases in Twitter data has been to 'convert' these data to more conventional social science data, for example, by developing methods to make demographic biases explicit and to create demographically representative subsamples of data (Sloan et al., 2015). Alternatively, it has been suggested that particular demographic biases might be harnessed to explore populations that are under-represented in other sources of research data, young men in epidemiological research, for instance. We have to be clear about this and not infer claims from a social media data set to the

Table 1. Methodological challenges along the data pipeline.

	Methodological considerations		
	Population	*Sample*	*Method of data construction*
Database	Storage design and method shape the types of information recorded about users	Historic data storage decisions and technical query limitations may shape what data are included in samples	Considerations of cost, performance and business requirements for data storage may shape what data are collected and stored and how
Server software	Determines who or what has access to the service and what information is required to set up an account	Server capacity may restrict data volume delivered; geographical location of server may affect data delivered	Operates data management (e.g. spam removal and moderation, load balancing) shaping what data are collected
API	APIs may not recognise all characters (languages) effectively or be available to all operating systems/software development toolkits	A variety of differently structured samples may be available	Defines the scope and volume of what data can be collected, stored and queried
Harvesting method	Harvesting methods construct different views of the populations. Web scraping may be more likely to access the population of currently active users, which could be different to the population accessed via historical searches using an API	Web scraping will by-pass ‘official’ data samples, offering data from a sample of web pages. This sample may be affected by the ‘filter bubble’ of the person accessing the web pages. Use of third-party data may introduce additional sampling effects	Different harvesting methods have access to different types of data about the population and sample
Client software	Different clients may generate different information about the population. On some platforms, you may know what client generated the content (this used to be the case on Twitter), on many though you cannot know this	Some clients (apps) may receive more data than others (if harvesting through a client)	Different clients may produce distinctive forms of data and metadata, e.g., some may add geographic data by default and some might link directly to shared or re-shared material
Subject	Different subjects – human/non-human, demographically distinct – may characterise particular platform populations	User activities may shape sampling methods (e.g. official samples may focus on central or highly active users)	User practices and meanings shape the data generated and the claims that can be made from these

API: Application Programming Interface.

general population without careful methodological controls or infer claims about all users of a particular social media platform from a sub-sample of data unless similar steps to match the sample with the wider population can be taken. However, we should also recognise that *a priori* demographic categories (sex, ethnicity, social class or age, for instance) may not the most important variables for working with these data, for example, we might be interested in the ebb and flow of public debate over time or between different types of social media account (corporate news accounts, political parties and individuals, for instance). Or we might want to see how emergent social networks produce collectivities based on online activities, rather than reflecting external demographic characteristics. In short, the approach depends on the question that is being asked and the claims that we want to make.

Third, and finally, we must consider what these data are, what they can tell us and what they cannot. We have already suggested that functionalities cannot be conflated with human meaning or relationships: likes, re-posts, friends – may be indicative but are, in the end, designed functionalities of commercial data companies. To describe the patterns that they produce, it may be more appropriate to refer to activity, information flows and networks, without making claims about their social significance. Making social claims about social media data will be more robust if we draw on 'wide data' – that is, multiple sources of digital data – and, as some other social media data researchers increasingly coming to conclude (e.g. Freelon and Karpf, 2015; Hall et al., 2016), if we employ mixed methods, to include offline as well as online data, qualitative as well as quantitative information. Indeed, it may be in this assemblage of multiple data sources, harnessed by theoretical understanding and methodological clarity, that social media data find their most powerful contribution for understanding the social world.

Funding

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

Notes

1. *Cyberpsychology, Behavior, and Social Networking*, *Journal of Social Networks* and *Journal of Social Network Analysis and Mining*.
2. *Psychological Science*, *Journal of Business Ethics*, *New Media and Society*, *Politics*, *Marketing Science*, *Public Administration Review*, *Social Indicators Research*, *Geoforum*, *Journal of Communication*, *Environment and Planning A*, *Journal of Health Communication*, *Health Policy*, *Annals of the Association of American Geographers* and *American Behavioural Scientist*.
3. http://www.internetworldstats.com/stats.htm (accessed 10 August 2017).
4. http://digital-stats.blogspot.co.uk/2012/07/demographics-of-uk-twitter-users.html (accessed 5 October 2016).
5. http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/ (accessed 5 October 2016).
6. http://www.statista.com/statistics/290447/age-distribution-of-us-whatsapp-users/ (accessed 5 October 2016).
7. Suggesting that the 2% using this function are more active than average, producing 3% of content.

8. http://blog.falcondai.com/2013/06/666-and-how-twitter-samples-tweets-in.html (accessed 5 October 2016).
9. We collected data from both the Search Application Programming Interface (API) and also data (at 1%, 10% and 100% of the Twitter Firehose) from an official Twitter data reseller; both queries were based on a specific set of hashtags, during the same time period. We found that the Search API contained significantly less 'Retweet' statuses compared to the data obtained from the official data provider, similar to the figures stated by Gonzalez-Bailon et al. (2012).
10. Note that some papers used more than one method of data harvesting.
11. https://unionmetrics.zendesk.com/hc/en-us/articles/201201636-What-do-you-mean-by-Twitter-reach-exposure-and-impressions- (accessed 5 October 2016).
12. 'Veien blir til mens du går'.

References

Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. *Wired*. Available at: http://www.wired.com/2008/06/pb-theory/ (accessed 23 June 2008).

Bowker G (2005) *Memory Practices in the Sciences*. Cambridge, MA: The MIT Press.

Bowker G (2013) Data Flakes: an afterword to raw data is an oxymoron. In: Gitelman L (ed.) '*Raw Data' Is an Oxymoron*. Cambridge, MA: The MIT Press, pp. 167–172.

Bowker J and Starr SL (1999) *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: The MIT Press.

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 (HICSS),* Honolulu, HI, 5–8 January, pp. 1–10. Los Alamitos, CA: IEEE Computer Society.

Bruns A and Stieglitz S (2012) Towards a more systematic Twitter analysis: metrics for tweeting activities. *International Journal of Social Research Methodology* 16(2): 91–108.

Burgess J and Bruns A (2012) Twitter archives and the challenges of big social data for media and communication research. *M/C*. Available at: http://journal.media-culture.org.au/index.php/mcjournal/article/view/561

Burgess J and Bruns A (2015) Easy data, hard data: the politics and pragmatics of Twitter research after the computational turn. In: Langlois G, Redden J and Elmer G (eds) *Compromised Data: From Social Media to Big Data*. London: Bloomsbury, pp. 93–111.

Busher T (2013) Objects of intense feeling: the case of the Twitter API'. *Computational Culture*. Available at: http://computationalculture.net/article/objects-of-intense-feeling-the-case-of-the-twitter-api

Chu Z, Gianvecchio S, Wang H, et al. (2010) Who is tweeting on Twitter: human, bot, or cyborg?'. In: *Proceedings of the 26th annual computer security applications conference.* Austin, TX, 6–10 December.

Chu Z, Widjaja I and Wang H (2012) Detecting social spam campaigns on Twitter. In: Bao F, Samarati P and Zhou J (eds) *International Conference on Applied Cryptography and Network Security.* Berlin, Heidelberg: Springer, pp. 455–472.

Doré B, Ort L, Braverman O, et al. (2015) Sadness shifts to anxiety over time and distance from the national tragedy in Newtown, Connecticut. *Psychological Science* 27(4): 363–373.

Driscoll K and Walker S (2014) Working within a black box: transparency in the collection and production of big Twitter data. *International Journal of Communication* 8: 1745–1764.

Evans L and Saker M (2017) *Location-based Social Media: Space, Time and Identity*. Basingstoke: Palgrave Macmillan.

Freelon D and Karpf D (2015) Of big birds and bayonets: hybrid Twitter in the 2012 presidential debates. *Information, Communication & Society* 18: 4390–4406.

Gaffney D and C Puschmann (2013) Data collection on Twitter. In: Weller K, Bruns A, Burgess J, Mahrt M and Puschmann C (eds) *Twitter and Society*. New York, NY: Peter Lang Publishing, pp. 55-68.

Gaffney D (2010) #iranElection: quantifying online activism. In: *Proceedings of the Websci10: Extending the frontiers of society on-line*, Raleigh, NC, 26–27 April.

Gehl R (2015) Critical reverse engineering. In: Langlois G, Redden J and Elmer G (eds) *Compromised Data: From Social Media to Big Data*. London: Bloomsbury, pp. 147–170.

Gitelman L and Jackson V (2013) Introduction. In: Gitelman L (ed.) *'Raw Data' Is an Oxymoron*. Cambridge, MA: The MIT Press, pp. 1–14.

Goldthorpe J (2016) *Sociology as a Population Science*. Cambridge: Cambridge University Press.

Gonzalez-Bailon S, Wang N, Rivero A, et al. (2012) Accessing the bias of samples in large online networks. *Social Networks* 38: 16–27.

Hacking I (2006) *Kinds of People, Moving Targets*. *British Academy Lecture.* Available at: http://nurs7009philosophyofinquiry.weebly.com/uploads/6/0/4/0/6040397/hacking_20071.pdf (accessed 2 August 2016).

Hall M, Mazarakis A, Peters I, et al. (2016) Following user pathways: cross platform and mixed methods analysis in social media studies. *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems (CHI EA '16)*, pp. 3400–3407. New York: ACM. Available at: http://snm.ku.dk/english/research/sections/evolutionary_genomics/employees/?pure=en%2Fpublications%2Ffollowing-user-pathways-cross-platform-and-mixed-methods-analysis-in-social-media-studies(943513b4-659f-4c6f-bd8b-fc167062c588)%2Fexport.html

Hardaker C (2016) Misogyny: machines, and the media, or: how science should not be reported. Available at: http://wp.lancs.ac.uk/drclaireh/2016/05/27/misogyny-machines-and-the-media-or-how-science-should-not-be-reported/ (accessed 2 August 2016).

Heer E and Verdegem P (2015) What social media mean for audience studies: a multidimensional investigation of Twitter use during a current affairs TV programme. *Information, Communication & Society* 18(2): 221–234.

Highfield T (2016) News via Voldemort: parody accounts in topical discussions. *New Media & Society* 18(9): 2028–2045.

Highfield T, Harrington S and Bruns A (2013) Twitter as a technology for audiencing and fandom. *Information, Communication & Society* 16(3): 315–339.

Humby (2014) Available at: https://www.marketingweek.com/2015/11/19/dunnhumby-founder-clive-humby-customer-insights-should-be-based-on-passions-as-well-as-purchases/ (accessed 5 September 2016).

Kitchin R and Lauriault T (2013) Towards critical data studies: charting and unpacking data assemblages and their work. The Programmable City Working Paper no. 2. Maynooth University, Ireland. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2474112

Leetaru K, Wang S, Cau G, et al. (2013) Mapping the global Twitter heartbeat: the geography of Twitter. *First Monday* 16, 5 April. Available at: https://journals.uic.edu/ojs/index.php/fm/article/view/4366/3654

Manovich L (2001) *Software Takes Command*. London: Bloomsbury.

Marres N and Gerlitz C (2015) Renegotiating relations between digital social research, STS and the sociology of innovation. *Sociological Review* 641: 24–46.

Marres N and Weltevrade E (2013) Scraping the Social? Issues in live social research. *Journal of Cultural Economy* 63: 313–335.

Mayer- Schönberger V and Cukier K (2013) *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Meier F, Elsweiler D and Wilson M (2014) +More than liking and bookmarking? Towards understanding twitter favouriting behavior. In: *Proceedings of the international AAAI conference on web and social media*, North America, May. Available at: http://www.cs.nott.ac.uk/~pszmw/pubs/icwsm2014-favouriting.pdf

Morestatter F, Pfeffer J, Liu H, et al. (2013) Is the sample good enough? Comparing data from Twitter's streaming API with Twitter firehose. In: *Proceedings of the 7th international conference on weblog Fs and social Media (ICWSM'13),* Cambridge, MA, 8–3 July, pp. 400–408. Palo Alto, CA: The AAAI Press.

Pariser E (2012) *The Filter Bubble: What the Internet is Hiding from You*. New York: Penguin.

Patterson D and Hennesey J (1998) *Computer Organization and Design: The Hardware/software Interface*. 2ne ed. San Francisco, CA: Morgan Kaufman.

Rodríguez-Amat J and Brantne C (2016) Space and place matters: a tool for the analysis of geolocated and mapped protests. New *Media & Society* 18: 1027–1046.

Ruppert E, Law J and Savage M (2013) Reassembling social science methods: the challenge of digital devices. *Theory, Culture & Society* 30: 422–446.

Sloan L, Morgan J, Burnap P, et al. (2015) Who Tweets? Deriving demographic information from Twitter' in meta-data. *PLoS ONE* 10(3): e0115545.

Wang Y, Callan J and Zheng B (2015) Should we use the sample? Analysing datasets sampled from Twitter's stream API. *ACM Transactions on the Web* 9: 13.

Watts D (2011) *Everything is Obvious: How Common Sense Fails*. London: Atlantic Books.

Zhang Y and Lu J (2016) Discover millions of fake followers in Weibo. *Social Network Analysis and Mining* 6(1): 1–15.

Zhao L, Lu Y and Gupta S (2012) Disclosure intention of location-related information in location-based social network services. *International Journal of Electronic Commerce* 16(4): 53–90.

Author biographies

Susan Halford is a professor of Sociology and a Director of the Web Science Institute at the University of Southampton. In this context, her research interests centre on the epistemology politics of digital data, methods and artefacts.

Mark Weal is an associate professor of Web Science at the University of Southampton and director of the Centre for Doctoral Training in Web Science and Innovation. Mark's research interests include the application of Semantic Web technologies to pervasive systems, methodologies for the use of social media data, and online behavioural interventions. He is co-director of the LifeGuide programme of research, developing online digital public health interventions.

Ramine Tinati is a senior data scientist on the Microsoft Global Black Belt Team in Singapore. Prior to this, Ramine was a new frontiers fellow and senior research fellow in the Web and Internet Science group at the University of Southampton, UK, where he worked on the EPSRC funded project, SOCIAM, which involves developing methods and analytics to understand the development and connectivity of the Web. Ramine's interests lie in real-time big data stream processing and querying.

Les Carr is a professor of Web Science, head of the Web and Internet Science research group and the director of the Web Science Institute at the University of Southampton. He researches the impact of network technologies on our lives and economy, and in particular on the research and knowledge industries. Professor Carr runs "EPrints Services," a spinout that supports Open Access and Open Data repositories for scientific information and is programme leader for Masters courses in Web Technology and Web Science.

Catherine Pope is a professor of Medical Sociology and is a member of the Southampton Web Science Institute. Her current research focuses on health services and the use of digital technologies in care delivery. She supervises doctoral students in the Southampton Web Science Centre for Doctoral Training doing research about the Web.