Geoforum 52 (2014) 167–179

Contents lists available at ScienceDirect

Geoforum

journal homepage: www.elsevier.com/locate/geoforum

Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'

Taylor Shelton [a,*], Ate Poorthuis [b], Mark Graham [c], Matthew Zook [b]

[a] Clark University, Graduate School of Geography, 950 Main Street, Worcester, MA 01610, USA
[b] University of Kentucky, Department of Geography, 817 Patterson Office Tower, Lexington, KY 40506, USA
[c] University of Oxford, Oxford Internet Institute, 1 St Giles, Oxford OX1 3JS, United Kingdom

ARTICLE INFO

Article history:
Received 12 August 2013
Received in revised form 6 January 2014

Keywords:
Big data
Geoweb
Hurricane Sandy
Mixed methods
Socio-spatial theory
Twitter

ABSTRACT

Digital social data are now practically ubiquitous, with increasingly large and interconnected databases leading researchers, politicians, and the private sector to focus on how such 'big data' can allow potentially unprecedented insights into our world. This paper investigates Twitter activity in the wake of Hurricane Sandy in order to demonstrate the complex relationship between the material world and its digital representations. Through documenting the various spatial patterns of Sandy-related tweeting both within the New York metropolitan region and across the United States, we make a series of broader conceptual and methodological interventions into the nascent geographic literature on big data. Rather than focus on how these massive databases are causing necessary and irreversible shifts in the ways that knowledge is produced, we instead find it more productive to ask how small subsets of big data, especially georeferenced social media information scraped from the internet, can reveal the geographies of a range of social processes and practices. Utilizing both qualitative and quantitative methods, we can uncover broad spatial patterns within this data, as well as understand how this data reflects the lived experiences of the people creating it. We also seek to fill a conceptual lacuna in studies of user-generated geographic information, which have often avoided any explicit theorizing of sociospatial relations, by employing Jessop et al.'s TPSN framework. Through these interventions, we demonstrate that any analysis of user-generated geographic information must take into account the existence of more complex spatialities than the relatively simple spatial ontology implied by latitude and longitude coordinates.

1. Introduction

Digital social data are now practically ubiquitous. This data is nowhere more visible than on the Internet, as over two and a half billion people currently both actively produce content, and leave behind all manner of transactional records, from comments and 'likes' on Facebook to the different products one has viewed and purchased on Amazon. In addition to online traces, people, buildings, roads, machines, plants and animals, alike, are increasingly augmented with sensors and software algorithms that produce electronic records of all manner of social, economic, political and environmental processes. These sources of digital data combine to create what we call '*data shadows*' (Zook et al., 2013; Graham, 2013; Graham and Shelton, 2013), or the imperfect representations of the world derived from the digital mediation of everyday life. As these datasets grow exponentially, researchers, politicians, and the private sector have begun to focus on how 'big data' might allow potentially unprecedented insights into our world (Hey and Trefethen, 2003; Anderson, 2008; Floridi, 2012).

Much of the 'big data' being produced online through social media has a significant amount of geographic information attached to it, often in the form of latitude and longitude coordinates known as 'geotags', which provide the means for new ways of doing, creating, making, and enacting geography. This process of attaching geographic coordinates to user derived digital content – often referred to as the geoweb – means that big data shadows are intimately connected to the material lived geographies from which they were produced. As such, social media has evolved beyond a simple online repository of conversations, networked interactions, and sites for the consumption of media, and is instead a dynamic record of when and how we move through and act in space, linked to other individuals and actions co-existing with us in those spaces. It is this connection between the geographies of online big data and the material processes they represent, and in turn impact, that we interrogate in this paper. In other words, what can big data from geographically referenced social media reveal about

* Corresponding author.
E-mail addresses: jshelton@clarku.edu (T. Shelton), atepoorthuis@uky.edu (A. Poorthuis), mark.graham@oii.ox.ac.uk (M. Graham), zook@uky.edu (M. Zook).

http://dx.doi.org/10.1016/j.geoforum.2014.01.006
0016-7185/

material processes and practices? And what can our pre-existing knowledge about such material processes and practices tell us about the underlying spatialities of big data?

In order to call attention to the interrelations between the material world and its connections to the virtual practices of what might more accurately be called '*geo*social media', we highlight a case study of Twitter activity in the wake of Hurricane Sandy, which struck the eastern seaboard of the United States in late October 2012. The second-most costly storm in US history behind only Hurricane Katrina, Sandy wreaked havoc on New York City's infrastructural systems, creating iconic images of flooded subway tunnels and roadways, dangling construction cranes and a blacked-out Lower Manhattan. In spite of these disruptions, the material effects of Sandy on New York City and the lives of people living in affected areas were clearly reflected in their online social media activities, as well as in the online activities of people living thousands of miles away. As such, the hurricane offers an accessible way to describe the variety of sociospatial relationships embodied in these big data shadows.

This paper argues that Hurricane Sandy offers a useful lens for understanding the digital data shadows produced by intensely material phenomena. Applications of big geosocial media data are increasing common throughout a range of activities beyond just disaster response, from urban planning to market research to political activism, and this case study provides the basis for a series of broad methodological and theoretical interventions into research on big data and user-generated geographic information. Methodologically speaking, rather than simply focusing on how massive databases are causing necessary and irreversible shifts in social practices or producing unprecedented insights into the world around us, we instead argue that it is more productive to analyze how small subsets of big data, especially georeferenced social media information, can reveal a broader range of social, economic, political, and even environmental geographies. Utilizing a mix of qualitative and quantitative methods, we uncover both broad spatial patterns within this data, as well as understand how these data reflect the lived experiences of the people who are creating it. Conceptually, we seek to fill a gap in previous studies of the geoweb, which have often avoided explicitly theorizing the nature of sociospatial relations. Building on Jessop et al.'s (2008) Territory-Place-Scale-Network (or TPSN) framework for understanding sociospatial relationships, we analyze the territorial, platial, scalar and networked dimensions of digital data shadows to highlight the polymorphous and complex spatialities of user-generated content. This allows for a greater consideration of the relational geographies of big data and geosocial media, which have largely been neglected in the literature to this point, while retaining an attention to more conventional ways of understanding the spatialities of this data.

In the following sections, we first review the relevant literature, focusing on conceptualizations and problematizations of big data. We then turn to understanding how big and user-generated data sources have been utilized in disaster response situations, before discussing the potential for new theorizations of sociospatial relations in studies of the geoweb. This is followed by a discussion of our data collection and methods, with attention to the potentials of using geotagged tweets for social and spatial analysis. In the penultimate section, we turn to the case of Hurricane Sandy and use a series of cartographic visualizations to highlight the variegated and polymorphous nature of sociospatial relations represented by Sandy's data shadows. Finally, we discuss the possibilities for and limitations of future studies of big data shadows.

2. Contextualizing 'big data' and geosocial media

This work is framed within the context of an important shift occurring in the social sciences: the emergence of 'big data', or what has been referred to as the 'fourth paradigm' of scientific research (Hey et al., 2009; Mayer-Schonberger and Cukier, 2013). Big data's proliferation throughout the popular press as a buzzword comes with many different definitions, and it is important to recognize that it refers not just to a quantitative increase in the size of the datasets being analyzed, but also qualitative shifts in the ways we approach the study of society (boyd and Crawford, 2012). These shifts include an increase in the scope of the data being collected, the speed at and timeframe within which it is collected, and the notion that otherwise unrelated datasets might be cross-referenced and analyzed to produce some meaningful insight (Kitchin, 2013).

Perhaps the most prominent proponent of this new data-driven science has been Chris Anderson, the former editor of *Wired Magazine*, who sees the proliferation and availability of these new datasets as a way to generate more insightful, useful, accurate, or true results than more conventional specialists or domain experts who carefully develop hypotheses and research strategies in order to understand a given phenomena – heralding 'the end of theory' (Anderson, 2008). Anderson's notion has entered not only the popular imagination, but also the research practices of corporations, states, journalists and academics (Lazer et al., 2009; Leetaru, 2011; Issenberg, 2012; Lohr, 2012; see also Torrens, 2010 for a geographic perspective), driven by the idea that the data shadows of people, machines, commodities, and even nature, can reveal difficult-to-understand social processes, simply by applying sufficient computing power to these massive amounts of data. In other words, researchers no longer need to speculate and hypothesise; they simply need to possess enough data and allow algorithms to lead them to important patterns and trends in social, economic, political, and environmental relationships.

This kind of naïve technological determinism echoes a similar argument made a decade earlier about the so-called 'death of distance' (Cairncross, 1997) brought by the internet, which itself stimulated a range of more nuanced theoretical and empirical works on the geography of the internet. Anderson's hyperbole around the end of theory has also given rise to a range of critical responses from social scientists of all types. This critical approach to big data has been especially pronounced amongst those scholars studying the geographic contours of user-generated internet content, as notions of big data frequently incorporate elements of what have variously been called the geoweb or volunteered geographic information (Goodchild, 2007; Elwood, 2008; Elwood et al., 2012). Nonetheless, scholars are just now beginning to employ social media data to ask substantive questions about the geographies of production, use and consumption of big data (Takhteyev et al., 2012; Graham et al., 2013; Tsou and Leitner, 2013).

Two primary criticisms of such big data analyses have been their failure to attend to persistent methodological issues and their overblown claims to be able to deduce significant meaning out of data without relying on pre-existing theoretical frameworks. In arguably the most visible critique of big data so far, danah boyd and Kate Crawford note that "Big Data and whole data are also not the same" (boyd and Crawford, 2012: 669). Similarly, Muki Haklay (2012) has warned that too often, analysis of big social media datasets tends to privilege the perspectives of so-called 'outliers', rather than incorporating a representative sample of the population. So while big data can capture a whole host of social processes that were previously difficult to study because of their transactional nature,[1] it

[1] Transactional data is used to refer to data describing events, which until recently were not readily accessible. This could quite literally include data describing a financial transaction or purchase at a store, or more loosely the kind of social media data we discuss in this paper. Of course, for this kind of data to become useful when cross-referenced with other databases, these transactions must be digital and automatically registered, which, for instance, would tend to exclude individuals whose economic activities are predominantly informal or cash-based. It would similarly exclude anyone who chooses not to participate in social media or other similar services.