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.

remains partial and biased in important ways that should qualify any findings from such research (Manovich, 2011; Ruppert et al., 2013).

Meanwhile, proponents of big data have also been critiqued for their relatively naïve claims to have transcended the need for any domain expertise in the subjects they analyze (Graham, 2012). For example, the physicist Geoffrey West has supposedly 'solved the city' using mathematical approaches, without having ever read any work in urban studies (Lehrer, 2010), while others have used similar databases of Twitter activity to revive the 'death of distance' thesis (Leetaru et al., 2013). It should be noted, however, that others, especially geographers, have been a good bit more cautious. For instance, Miller (2010) argues that data-driven approaches have much to benefit from the inclusion of more conventional domain expertise, while more traditional approaches to social science can benefit from the improved hypothesis generating capabilities of data mining.

2.1. Social media and crowdsourcing disaster response

One realm in which the role of big and user-generated data has generated massive amounts of attention has been disaster response (cf. Goodchild and Glennon, 2010; Li and Goodchild, 2010; Liu and Palen, 2010; Roche et al., 2013). While key players in this space, such as Ushahidi and the crisis mapping community, developed in a more-or-less ad hoc manner in order to respond to disasters such as the 2010 Haitian earthquake, more established institutions, including states and international NGOs, are similarly promoting the potentials of these new data sources and their analysis for responding to disaster or crisis situations. For example, The Red Cross has actively been collecting tweets about disaster situations (Red Cross, n.d.), while the United States Geological Survey has been using its 'Did You Feel It?' online reporting tool to crowdsource reports about the intensity of earthquakes for over a decade (Wald et al., 1999; Wald and Dewey, 2005). Other less systematic examples include the much-publicized efforts of Newark, New Jersey mayor Cory Booker to personally respond to unfulfilled service requests during a major blizzard in that city, following attempts by residents to use Twitter to encourage a response (Gregory, 2010).

While there are many such examples of success in harnessing this kind of data for disaster response, two important contributions to this discussion from a geographic perspective raise questions about their efficacy. Crutcher and Zook (2009) and Zook et al. (2010), discussing the use of social media in response to Hurricane Katrina and the Haitian earthquake respectively, argue that patterns of adoption and utilization of such technologies in disaster response have largely followed long-standing patterns of sociospatial inequality, producing uneven data shadows that do not reflect the on-the-ground realities following disasters. This is driven, or at least exacerbated, by the fact that such disasters typically represent the failures or inadequacies of state-based disaster relief, leading to a greater number of citizens taking an active role in the production of information about such events (cf. Leszczynski, 2012, on the relationship between the rollback of state functions and the production of geographic information).

In the case of Hurricane Katrina, Crutcher and Zook were able to show that the production of user-generated, geotagged reports tended to be associated with wealthier, whiter, more tourist-oriented locations within New Orleans, despite the greatest effects of the storm being felt in predominantly poor and black areas, such as the Lower Ninth Ward. As the later case of the Haitian earthquake of 2010 demonstrates, however, such disasters *can* serve to stimulate greater attention to the production of user-generated geographic information in and about such marginalized places. Indeed, some of the most striking examples of the volunteer effort following the earthquake are those that demonstrate the lack of codified and widely-accessible geographic information about the country prior to the earthquake, and the explosion of information produced following it in order to aid in the recovery effort (cf. Zook et al., 2010 for visualizations of the growth in user-generated Google placemarks following the earthquake, or ITO World, 2010 for a time-lapse video of edits to OpenStreetMap). Nonetheless, these findings demonstrate that while such participatory, citizen-driven and technology-centric efforts have great potential to aid in disaster situations, these solutions are only ever partial, both in terms of participation and assistance, and are no replacement for more coordinated 'on the ground' relief efforts.

As such, the case study of Hurricane Sandy used in this paper represents an important opportunity to revisit these earlier findings in a different context. Apart from the contextual differences between New York City and New Orleans or Port-au-Prince, one major difference between Sandy and the earlier cases hinges on technology. While Google Earth had just been released when Hurricane Katrina struck the Gulf Coast in 2005, and the Haitian earthquake represented something of a test case for technology-based disaster response at a distance, the nearly 20 million tweets about Hurricane Sandy (Twitter, n.d.) provide a sufficiently robust source of data to map the data shadows of the storm. This wealth of user-generated data can help us in better understanding the connections between the material world and its virtual representations. It also allows us to articulate a more coherent conceptualization of the spatialities of these data shadows in order to counter the dominant popular discourse that sees big data as an objective and normatively superior way of understanding the world, and to fill conceptual gaps that remain in the critical literature on these issues.

But as Kate Crawford (2013) points out, referencing both the case of Hurricane Sandy and the aforementioned paper by Crutcher and Zook, one cannot rely solely on social media content to reveal where the most damage occurred. Just because there is more data from which to work does not mean the aforementioned problems of representation and unequal power relations embodied in the data are resolved. The intimate intermingling of digital and material facets of life means that the production of geosocial media content is often strongly connected to place-based features and events, but also that longstanding inequalities and situational or contextual constraints distort the representativeness of such data sources. While we are sympathetic to such critiques of big data, we maintain that an explicitly geographical approach might be able to partially resolve some problems raised by earlier critiques. For example, while using geotagged tweets as one's sole data source might produce a flawed or incredibly partial analysis of an event like Hurricane Sandy, this data can also be used to answer broader questions around the geographies of the geoweb and how such spatialities might be conceptualized, as we do in this paper.

2.2. The polymorphous geographies of social media

Research into the geographies of social media has largely eschewed any explicit theorization of space and spatiality. Even where implicit, studies have tended to privilege a unitary understanding of space. For example, Takhteyev et al. (2012) employ a networked or topological understanding of sociospatial relations by focusing on social connections between Twitter users, while work by Goodchild and Li (2012); Haklay (2010) focused on questions about the quality and locational accuracy of volunteered geographic information. Other similar work on mapping the user-generated and social media data from the geoweb has alternatively tended to over-emphasize the groundedness of such content in particular places, or how particular place-specific attributes, such as religion and language, are reflected in this data (Graham and Zook, 2011; Shelton et al., 2012; Graham and Zook, 2013).

Yet conceptualizations of space that focus on only a single understanding of it necessarily belie the complexity of forms that sociospatial relations take. In order to overcome this issue, Crampton et al. (2013) have proposed a loose framework for going 'beyond the geotag' in analyzing geosocial media data. They argue that researchers should explicitly recognize the diversity of spatialities embodied in social media content in order to avoid over-privileging what amounts to a simplified spatial ontology of latitude and longitude coordinates. Analyses that fail to go beyond a simplified spatial ontology – e.g. simply plotting data points in Cartesian space – often overlook the range of quantitative and qualitative approaches that allow one to better understand the context and meaning of such big data, and tend to reinforce territorial or place-based dimensions of data at the expense of thinking space relationally (cf. Massey, 1991; Amin, 2002).

We use this constructive critique of earlier work on mapping user-generated data as a foundation for positioning our intervention within a pre-existing framework for understanding sociospatial relations. Specifically, we adapt Jessop et al.'s (2008) TPSN framework in order to construct a more holistic picture of the variegated landscapes of the geoweb, emphasizing both the territorial and relational dimensions of this data. Jessop et al. argue that by focusing on the polymorphous nature of sociospatial relations and their expression through the dimensions of territory, place, scale and networks, a more open and comprehensive understanding of sociospatial relations is possible. They note that most sociospatial research is concerned with just one of these dimensions, committing what they refer to as the fallacy of 'one-dimensionalism'.

Instead, Jessop et al. offer the TPSN framework as a kind of metatheory to emphasize the complex and variable nature of sociospatial relations as simultaneously and variably bounded and coherent (territory), as differentiated and embedded in particular contexts (place), as hierarchically organized (scale) and as interconnected or interdependent (networks). Each of these dimensions must be understood as always co-present and interconnected with the others; they can only be separated analytically, but never in practice, as, for example even the most global phenomena are always grounded in particular experiences of place, and vice versa. This approach avoids privileging any single dimension of space and instead highlights the ways that the technologies and knowledges of the geoweb and social media are expressed in a number of different ways simultaneously. It is for this reason that our empirical analysis, taken up in Section 4, does not separate out each of these dimensions when considering different ways of looking at the data. While some dimensions are more prevalent in a given representation than others, no representation is illustrative of just a single dimension.

As the TPSN theoretical framework is a key part of our analysis, further explanation of each of the four dimensions of sociospatial relations is warranted. While some of the specifics of Jessop et al.'s explanation of TPSN may not be especially relevant in the case of Hurricane Sandy (for instance, this paper does not focus on divisions of labor or nongovernmental international regimes which are important in the context of political-economic analysis that the framework was originally intended for) their framework offers a useful heuristic for thinking about the multiple spatialities of social media data, or any social phenomenon for that matter. Table 1 outlines their original conceptions of each of the four dimensions, as well as our adaptations of these ideas to the context of analyzing Hurricane Sandy's data shadows.

We employ the concept of *territory* to understand how user-generated content is spatialized in particular localities through the mirroring of offline, material phenomena occurring there. While resembling a conventional definition of the spatiality of big data as simply a set of latitude and longitude coordinates, it more importantly provides insight on the general contours, and occasional discrepancies, between our understandings of the materiality of a given phenomena and its online reflection. Such a connection to particular localities is tempered by integrating a focus on *scale*. While scale is a slippery concept – varying in meaning depending on whether one is using the concept in the context of an urban political economy, physical geography or GIScience approach (just to name a few competing understandings), our attention rests on the ways that using different scalar constructs, such as the juridical boundaries of neighborhoods, zip codes, census tracts, cities, and states, can alter perceptions of the sociospatial processes embodied in these data shadows.

In addition to territory and scale we also integrate a focus on *place*, or the lived dimension expressed in the qualitative information contained within these datasets. Rather than assuming a simple relationship between a piece of social media content and the location to which it is tagged, we work to understand the significance of these localities to the users producing such representations and the social contexts in which such content is embedded, acknowledging the potential for these experiences to be spatially distanciated from the locations in which a given event occurs. For this reason, we shift the notion or theme of 'proximity' from place, as it is conceived by Jessop et al., to territory, as mentioned above, to preserve an understanding of place that is more closely aligned with conventional understandings within geographic thought (Cresswell, 2004). Finally, we turn to connecting these lived experiences of place to the broader patterns evident in territorial and scalar frameworks, through a focus on sociospatial *networks*, or relational spaces. That is, understanding territories, places and scales as bounded or limited ignores the connections between localities, and the ways that social processes are increasingly extensive over long distances. In short, the network dimension reflects that one cannot fully understand the geographies of place-based phenomena without understanding that place's connections to other localities.

Rather than simply gathering such data, aggregating them and then displaying their location on a map, the TPSN approach provides a richer set of sociospatial dimensions that can be used to understand the production and consumption of geographically referenced big data such as that which is derived from social media. Utilizing the TPSN framework also provides an important connection between research on the geoweb and big data to the broader field of geography and sociospatial theory. It allows us to clearly demonstrate that the sociospatial relations of geosocial media are not divorced from sociospatial relations more generally, and ultimately helps illuminate the full range of human experiences that are evident in such data shadows.

3. Collecting and analyzing big data from social media

In order to operationalize the TPSN framework in the context of big data, this paper analyzes the data shadows of Hurricane Sandy through a specially designed software program that collects all geocoded tweets worldwide through the Twitter API, or application programming interface. While websites and social media platforms often provide APIs as part of a business strategy, researchers have begun to take advantage of these tools to access the significant amounts of data being generated through these platforms. The specially designed software used here was already operating prior to Hurricane Sandy, allowing us to select only tweets sent from within the United States between October 24 and October 31, 2012 that contain the keywords 'sandy', 'frankenstorm',[2] 'flood',

[2] The term "Frankenstorm" was widely used to refer to the landfall of Hurricane Sandy in the northeastern United States in late October 2012. The term was adopted both because of the intensity of the storm and its timing immediately before Halloween.

Table 1
Operationalizing the TPSN framework.

Dimension	As articulated by Jessop et al. (2008)	Our operationalization
Territory	Bordering, bounding, parcelization, enclosure	Locality, proximity, materiality
Place	Proximity, spatial embedding, areal differentiation	Lived experiences, individual perceptions
Scale	Hierarchization, vertical differentiation	Hierarchical organization, 'size' of areal lens
Networks	Interconnectivity, interdependence, rhizomatic differentiation	Interconnectivity, non-proximate, relational space

or variations thereof. This results in a dataset consisting of 141,909 tweets. While each tweet has a variety of associated metadata, ranging from the actual tweet text to the number of friends that that Twitter user has, this study only uses the actual text, the timestamp and the location of the tweet.

It should be noted that these 141,909 geotagged tweets represent only a fairly small percentage of the total number of Sandy-related tweets during this time period, as only approximately 1.7% of all tweets contain explicit geographic information. While techniques exist to derive locational information from user-provided location information in profiles, this introduces its own set of issues surrounding self-reporting, precision, geocoding accuracy and the difference between a user's home location and the location from where a particular tweet is sent.[3] It is for this reason that we focus only on the relatively clean dataset of tweets that contain explicit geographic information. But even within this dataset, there exists variation in how location is derived. Of our dataset, 82% of the geotagged tweets contain an actual latitude/longitude coordinate pair, derived from the GPS sensor on a smartphone or through cell tower triangulation. The other 18% only contains a 'place' specification, which can vary in precision from the country level to cities to neighborhoods or points of interest. For obvious reasons, tweets that only have higher-level place information are filtered out when doing a local level analysis (e.g., tweets with only city-level definitions must be discarded when doing a neighborhood-level analysis).

Given the relatively large dataset – thousands of points – one must be mindful of three significant challenges, that if not dealt with correctly can prevent even the relatively straightforward exercise of mapping points in Cartesian space from yielding useful insights. First is the issue of overplotting. Plotting thousands of points onto a single map makes it difficult to distinguish between the intensity or size of different clusters. Second, regardless of the phenomena under study, places that are already large content producers will almost certainly produce high amounts (in absolute terms) of social media references to the phenomenon of interest. The third, and related, challenge is that the uneven spatial distributions of tweets means the amount located in any one region varies considerably, affecting the confidence with which we can infer differences from location to location.

To overcome these three challenges we use an approach that overlays the area of study with a grid of hexagonal cells of varying size. We use hexagonal cells instead of the more common rectangular grid cells for two specific reasons. First and foremost, cartographically, hexagons make it easier to increase the size of each cell (thus negating the use of smoothing, which is not always a good practice when dealing with phenomena that are not necessarily 'smooth') while still allowing the reader to discern contours. Square cells, as opposed to hexagons, are much more distracting to map readers and thus make it more difficult to determine the spatial pattern of a phenomenon (Carr et al., 1992). Second, hexagons also have a higher representational accuracy (Scott, 1985) and, when used in statistical analysis share a direct boundary with 6 neighbors, instead of the 4 direct neighbors of squares. Being able to vary the size of the cells allows us to use 'appropriate' cell sizes for different scale levels as well as address the potential effect of the Modifiable Areal Unit Problem (cf. Poorthuis, 2013, for a more detailed discussion of this approach). In this paper, we use 65-km wide cells for the national scale and 2-km wide cells for the urban scale – both chosen to balance the generic with the particular so the map reader can distinguish larger patterns while not losing some smaller idiosyncrasies. Furthermore, we use a sample of 138,021 random tweets sent from the United States during the same time period from which our database of Sandy-related tweets was drawn in order to normalize data within these hexagonal units. Although population is often used for normalization purposes, using a random sample of tweets allows us to normalize by 'Twitter population' instead. The sample is drawn from the same proprietary system as the Sandy dataset, which allows for the extraction of random samples of all geotagged tweets of any size. In this case, we have chosen the sample to be roughly the same size as the dataset under study. We calculate both the number of Sandy-related tweets as well as the number of 'random' tweets. We then use both counts to calculate a variation on the odds ratio, referred to as location quotient in spatial economics, taking the lower bound of the 99.9% confidence interval for each cell as follows:

$$OR_{lower} = e^{\ln(OR_i) - 3.29 * \sqrt{\frac{1}{p_i} + \frac{1}{p} + \frac{1}{r_i} + \frac{1}{r}}}$$

where p_i is the number of tweets in hexagon i related to the phenomenon of interest and p is the sum of all tweets related to the phenomenon; r_i is the number of random tweets in hexagon i and r the sum of all random tweets. This results in a ratio where a value of 1 means that there are exactly as many data points for the phenomenon as one would expect based on the random sample. An odds ratio greater than 1 means that we can say, with 99.9% confidence, that there are more points related to the phenomenon than one should expect, and vice versa for anything under 1.

Although the entire dataset contains more than 3 billion tweets as of August 2013, the case studies in this paper only use a subset of this data based on the query outlined previously. It is important to highlight that we cannot draw direct correlations between the size of our datasets and the veracity of insights that can be drawn from those data. Although these data offer the raw materials for analysis and understanding, simply plotting points on a map is an insufficient endeavor to comprehend the polymorphous and variegated geographies of social media as conceptualized using the TPSN framework. As such, we will augment a more quantitative and GIS-oriented analysis with a qualitative analysis of the content of tweets. Such analysis is not a significant departure from longstanding traditions of cultural landscape interpretation within geography, though the landscapes that we interpret here are the digital representations of material actions, patterns, and processes, or what have previously been referred to as 'cyberscapes' (Crutcher and Zook, 2009; Graham and Zook, 2011; Shelton et al., 2012). This paper's methodological approach is thus necessarily interlinked with the conceptual approach of the TPSN framework.

[3] See Stephens and Poorthuis (2013), who were able to find location data for 25% of all users, and Graham et al. (forthcoming), who show that geocoding accuracy varies substantially based on both location and language, for more discussion of these issues.

4. Sociospatial dimensions of Hurricane Sandy's data shadows

In order to better understand the diversity of ways that social media data shadows reveal or conceal useful information, we now turn to interrogating the aforementioned dataset of tweets related to Hurricane Sandy through the four core dimensions of sociospatial relations – territory, place, scale and networks – as outlined by Jessop et al. (2008). While each of the visualizations might be loosely placed under one of the four headings, we have intentionally chosen not to present them separately, so as to emphasize that each visualization demonstrates the fundamentally multiplicitous sociospatial relationships of the geoweb.

The first, and most obvious, way to approach these data is to look at the distribution of Sandy-related tweets at a broad spatial scale, in this case looking at the continental United States. Using the odds ratio metric explained in the previous section, Fig. 1 clearly shows a significant concentration of Sandy-related tweets along the eastern seaboard of the US, especially in those places that were most affected by the storm, with approximately 30% of all Sandy-related tweets being located in the New York City metropolitan area. While there are some intriguing anomalies, for instance the cluster of tweeting around Phoenix, Arizona, this map is largely unsurprising given the material manifestation of Hurricane Sandy in the Northeastern US.

Zooming into the affected area, there appears to be important utility in employing social media data to measure the extent of storm damage (see Fig. 2). Using the same data as Fig. 1, this map adds a layer representing the official 'High-Impact Zone' as determined by the Federal Emergency Management Authority (FEMA), which is roughly congruent with the areas with the highest relative amounts of Sandy-related tweeting activity. This connection is further bolstered by the fact that the New York metropolitan area suffered the greatest financial losses from the storm, totaling approximately $19 billion (Gormley, 2012).

To be clear, this map is not intended to discount that other populated areas, such parts of Pennsylvania, Virginia, and the Caribbean (which we have excluded altogether from this analysis) were also hit hard by Sandy. Rather, we use the 'High Impact Zone' definition in Fig. 2, to demonstrate a clear connection to the places in which that content was produced, underlining the territoriality of geosocial media data. But highlighting this territoriality is merely the first step of the analysis. In order to place the groundedness of this content in context, we must also examine how it is intertwined with other dimensions of sociospatial relations.

For example, despite the overall devastation experienced by New York City and the surrounding areas, it is problematic to assume that New York City as a place is entirely coherent and that people's experiences of the storm were uniform throughout different areas of the city. By integrating a focus on scale with our already established focus on territory, we can get a better idea of the actual contours of Sandy-related tweeting in New York City (see Fig. 3).

When taking a closer look at New York City, we can adjust the size of the hexagonal cells used to aggregate tweets, which in turn creates a finer grained surface for analysis. While we are still examining the territoriality of tweets, we have also in this moment shifted scales, essentially disaggregating the coarser definition of the New York metropolitan area used in Figs. 1 and 2 into a series of smaller spatial units to allow for intra-urban analysis. Fig. 3a highlights (via text call outs in the maps) places in the city where significant events during the storm coincide with higher-than-average levels of tweeting, while Fig. 3b highlights places where major events were reported by the media but had relatively few tweets.

Fig. 3a demonstrates that a number of places that experienced significant damage were also major producers of Sandy-related tweets. Some areas with significant tweeting activity, such as the Lower East Side, which experienced significant flooding and power outages, are relatively wealthy, and even some poorer areas, such as Coney Island, had significant levels of tweeting activity.

At the same time, however, some of the hardest hit places also had relatively little tweeting activity (see Fig. 3b). For example, in Breezy Point, a fire destroyed more than eighty homes, but only a handful of tweets come from that location. Sandy inflicted similar damage on large parts of the Rockaway Peninsula with very little mention in these places on Twitter. We are also able to see a general lack of tweeting from Staten Island, which has the unfortunate distinction of having nearly half of the Sandy-related deaths within the city, not to mention massive amounts of property damage in the Oakwood area. While some residents in these areas were likely preoccupied with more pressing matters than tweeting, this runs counter to examples in Fig. 3a where significant amounts of

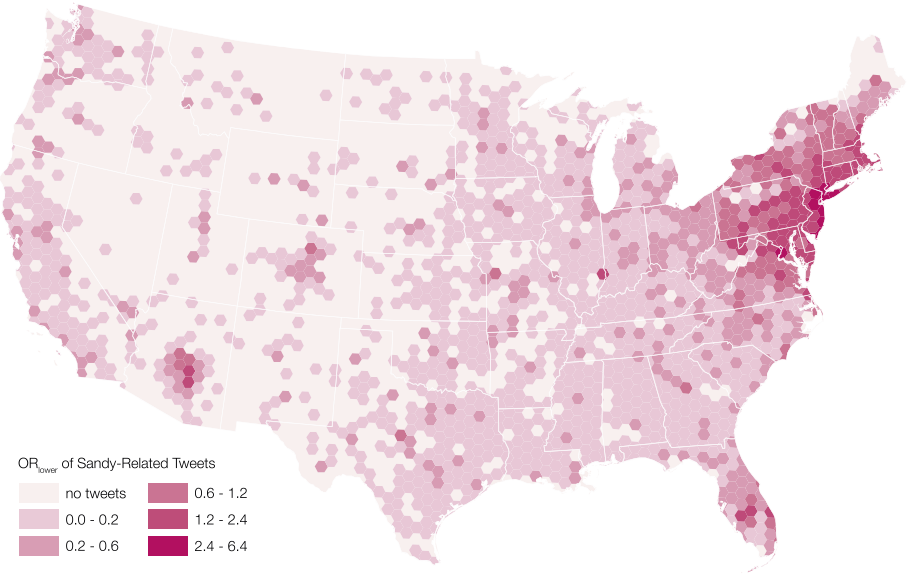

Fig. 1. Sandy-related tweets across the United States.

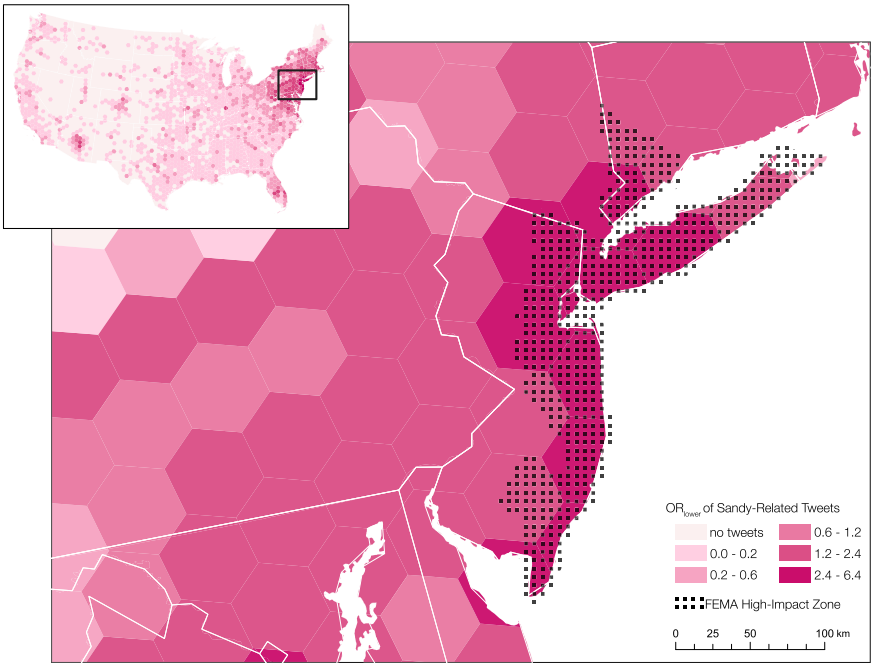

Fig. 2. Sandy-related tweets along eastern seaboard.

tweeting correlated with high-damage locations. The differences between these two figures suggest that places on the spatial periphery of the metropolitan area, e.g., Staten Island or the Bronx, are more likely to be marginalized within data shadows than more central locations, e.g., Manhattan and Brooklyn. While there is no definitive explanation for these discrepancies between damage and tweeting activity, it is above all demonstrative of the fact that the correlation between these variables changes across scales, thus necessitating the inclusion of the scalar dimension in any similar analysis.

Thus, shifting the analysis from the national to the urban scale reveals that the relatively strong correlation between tweet density and territories most affected by Sandy breaks down at finer scales of analysis; a finding that raises concerns about some of the practical applications of mapping geosocial media in disaster situations. In other words, strategies that rely upon the data shadows of social media for determining the allocation of scarce resources in a crisis need to consider the biases and permutations that accompany the production of this data.

For this reason we argue for the utility in proceeding with an iterative analysis that focuses on specific events, rather than simple mappings of terms like "sandy", "frankenstorm" and "flood." For instance, mapping the location of the 774 tweets mentioning "crane" during the storm, we are able to pinpoint the location of the now infamous 57th Street crane that was left dangling in the aftermath of the storm (see Fig. 4). Although we are cautious about the potentials of automated, algorithmic analysis of big data in many contexts, this example highlights the potential of such analysis in places characterised by thick data shadows, such that a kind 'early detection' mechanism might be able to automatically identify spatial and temporal irregularities in the data.

In addition to highlighting the territoriality of the event at the scale of the neighbourhood, a more in-depth examination of the actual content of the tweets reveals a diversity of opinions expressed about the crane mishap. These run the gamut from those who lived or worked in the area and were relaying their own personal observations, to those choosing to make light of the situation, rather than dwell on its potentially disastrous consequences for those in the immediate vicinity. For example, some Twitter users used simply exclamations like "Big scary broken crane!", often accompanied by photos, while others made joking commentary such as "Crane dangling tourism [is] better than regular tourism" or "How do you get to Carnegie Hall?. . .be a crane at a nearby construction site and wait for a hurricane to blow you there".

The 57th Street crane example demonstrates the value in extending our analysis beyond the territorial and scalar dimensions and into the lived dimension of place-making. As useful as it is to use these virtual expressions of material phenomena to locate these events in Cartesian space, stopping there neglects the way that these data are reflective of particular experiences of place by particular individuals. This is, we assert, a necessary, but thus far largely overlooked, contribution that geographers can make to the broader study of social media activity. A focus on the qualitative experience of place embodied in this data and resulting data shadows offers a significant opportunity for geographers and others interested in the spatial dimensions of social media, and can create a much more nuanced understanding of these dimensions when paired with the more general analysis of territory and scale emphasized in Figs. 1–3. While this may have more to do with the post hoc analysis of such catastrophes than for the immediate disaster response, it highlights the importance of attending to the qualitative information and social context of such data, even during disaster response, and not over-privileging automated systems for sentiment analysis, which leave significant potential for misinterpretation.

Another important consideration is that a focus merely on the greatest concentrations of tweeting activity provides relatively little insight into the array of meanings encoded into social media datasets. While we can use the first slice of the territorial dimension to understand the basic spatial distribution of tweeting activity (as evidenced in Figs. 1–3), this assumes a level of homogeneity within the qualitative information contained within the tweets themselves. It is similarly important to consider that places which may not have especially high concentrations of tweeting activity, and which might be quite far from those places which do, also have something to tell us about the spatiality of social media. For

Fig. 3. Sandy-related tweets in New York City metropolitan area.

instance, of the nearly 142,000-geotagged tweets used in Fig. 1, only 42,000 or so of those are in the New York metropolitan area. So what are we to make of the remaining 100,000 tweets if we focus only on those places with the most activity? Indeed, what is the utility of 'big data' if we are ignoring such a significant portion of it?

One corrective to this, inspired by Doreen Massey's idea of a global sense of place (Massey, 1991), is to turn our attention to a greater diversity of places, including those with relatively few Sandy-related tweets and those quite far from New York in absolute distance, but actually quite proximate in relational terms. By combining a focus on the place and network dimensions of sociospatial relations as outlined in the TPSN framework, we can begin to put a greater emphasis on understanding the totality of the dataset. For example, looking at Sandy-related tweets in the Los Angeles metropolitan area, of which there are only 2476, one sees a number of revealing inter- and intra-urban geographies (see Fig. 5). Although Los Angeles as a whole was thousands of miles away from the physical manifestations of Hurricane Sandy, the data shadows produced by

Fig. 4. Tweets about the 57th street crane in New York City.

Twitter users in different parts of this metropolitan area vary considerably from each other.

With New York City and Los Angeles being the archetypical 'global cities' of the United States, we know that one of the most important linkages between them is by way of air travel (Derudder et al., 2007), so a cluster of people in each city's airports concerned about their ability to fly cross-country with the impending storm is perhaps unsurprising. But a variety of relational connections are visible in other locales as well, even in the absence of similarly obvious rationales. For instance, though Hollywood has the greatest overall number of Sandy-related tweets in Los Angeles, tweets originating from this area seem filtered through a film and television-centric framing specific to the area, with references linking Sandy to films like *Judgement Day* and *The Perfect Storm*, the reality television show *The Jersey Shore* and The Weather Channel's tendency to dramatize weather events in order to promote their own programming. Read through the lens of TPSN, understanding Los Angeles' place within the broader landscape of Sandy-related tweeting, as well as its internal variegation, brings to the forefront the dimensions of place and networks, or the "presence of both the proximate and the remote at the same geographical level" (Amin, 2002: 389).

While Los Angeles does not necessarily have a particularly prominent place in the territoriality of Sandy-related tweeting at any scale, these examples highlight the utility of going beyond just looking at those areas with the highest concentrations of tweeting activity. Instead, a closer reading of social media content in a variety of locations reveals how spaces that might otherwise be neglected in such analyses still provide important insights into the geographies of big data. Rather than simply matching information mediated by social media platforms to spatial locations, significant meaning can also be drawn from the interconnectivity and interdependence of those data, raising the question of what the topology of connections between information producers and information itself tells us about these material phenomena. In other words, we can see not only where something happens in physical space, but how an event connects to other spaces both near and far through network ties. Multidimensional understandings of sociospatial processes are important across a range of issues, beyond just our readings of such processes through big data. For instance, natural disasters like Hurricane Sandy are conventionally understood as having very particular localized effects around the areas most affected, in this case the New York metropolitan region and the eastern seaboard. But any understanding of Sandy as being spatially delimited to these places would be lacking. Equally important is how those impacts were shaped but by more complex social forces that stretch beyond these localities but nevertheless structure experiences of and responses to such events, such as the genesis of climate change and its impacts on increasingly irregular and volatile weather patterns or the political-economic structures that cause predominantly poor and minority neighborhoods to be the most vulnerable to such disaster events (cf. Smith, 2006). But such logics also apply to the ways that disasters or other such events reach beyond these localities in their effects, which are further mediated by place-based experiences in those locations, as is demonstrated by the different experiences and interpretations of Hurricane Sandy within Los Angeles as seen in the Twitter data.

Although the relational connection between New York City and Los Angeles makes for a convenient example, it is less clear how sociospatial networks across the US were articulated through social media during Hurricane Sandy. If the tweeting from LAX during Sandy is indicative of a broader pattern induced by

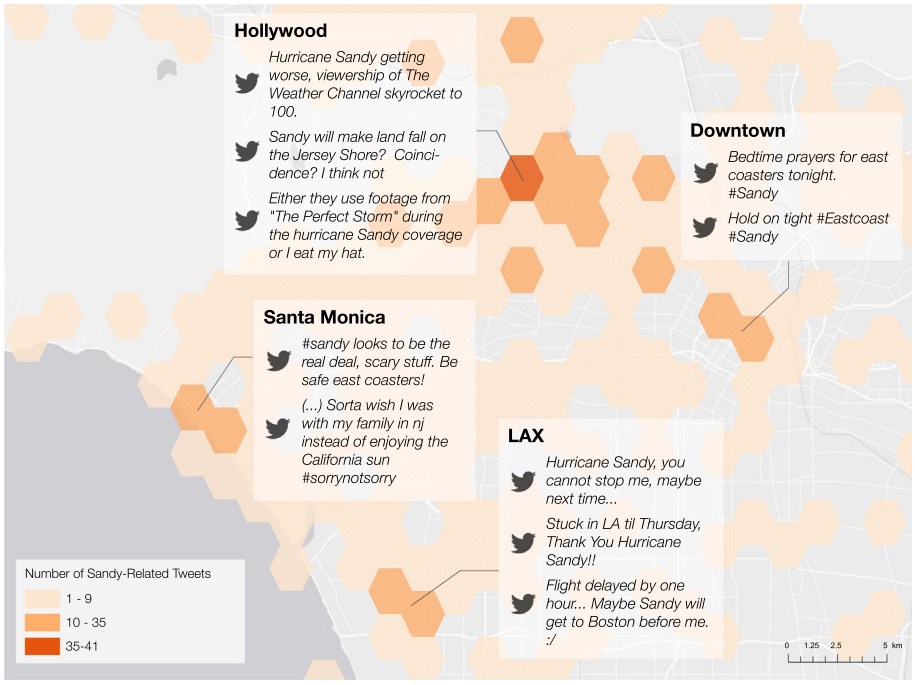

Fig. 5. Sandy-related tweets in the Los Angeles metropolitan area.

airplane-enabled connectivity, can we find similar connections to other locations across the US as well? The goal of such an analysis is to demonstrate the extent of the relational dimension of social media activity beyond the obvious connections between global cities such as New York and Los Angeles.

Using T-100 Domestic Market data from the Research and Innovative Technology Administration (RITA) on flights and the number of passengers between city pairs in 2012, we determined the 50 cities that have the most passenger traffic with New York City, ranging from Chicago (3.5 million passengers back and forth) to Kansas City (175,000 passengers). Since operations and activities at some airports close to New York were directly affected by Sandy's landfall, we exclude any airport within 500 km of Manhattan in this analysis. For the remaining airports we used a buffer of 5 km to collect all Hurricane Sandy related tweets and calculated the lower bound of the odds ratio as we did for the hexagonal cells used in Figs. 1–3. If relational networks did not play a significant role in Sandy-related tweeting, one would expect to see a direct distance decay effect: as the distance from New York City increases the odds ratio should decrease.

Fig. 6a, however, shows that physical distance has no significant relationship with the relative level of tweeting activity about Hurricane Sandy as is evidenced by both the scatterplot[4] and the map (Spearman's rho is −0.05). The map uses an azimuthal equidistant projection with New York City as the center, where the size of each airport is proportional to its odds ratio. Airports that are equally distant in physical terms from New York have widely diverging measures of Sandy-related Twitter activity. In addition, the average odds ratio in each 1000 km zone does not decrease the further away one travels from New York.

In contrast, Fig. 6b shows that the relational ties between each city and New York, measured by number of passengers, exhibits a much stronger positive correlation with the odds ratio metric of Twitter activity (Spearman's rho is 0.34). This figure preserves the directional bearing of each city with respect to New York City, but instead uses an inverse of the number of passengers to recalculate the relational distance between the cities. Airports are thus no longer displayed according to their physical distance from New York City, but rather based on the amount of passenger traffic between the two cities. Since the bearing has remained the same, airports with a higher intensity will move closer to New York along that line, and vice versa. In addition to the correlation coefficient, we can also visually determine that cities with a lower odds-ratio, such as Pittsburgh and Memphis, have a tendency to move towards the outer circles while cities with a higher odds ratio, such as San Francisco and Los Angeles, move relatively closer.

In other words, it is the relational connection to New York, measured by number of air travellers, not physical distance, which better explains the level of concern with Hurricane Sandy. This concern, however, can vary within metropolitan territories as evidenced by Fig. 5 depending upon the scale of analysis; some parts of an urban area may have much stronger relational ties to distant cities, while other parts are largely disconnected from such translocal flows.

To test the extent to which the data shadows of Sandy-related tweeting are a localized phenomenon within certain parts of metropolitan areas (rather than a more generalized territorial phenomenon), we increased the initial buffer around each airport from 5 km to 25 km. Thus, rather than just capturing neighborhoods that are spatially proximate to the airport, this measure captures a much wider swath of each metropolitan area. In the case of Los Angeles, this includes the entirety of the territory shown in Fig. 5 and beyond. With this larger buffer, there is a near-reversal of the correlations illustrated in Fig. 6, as Pearson's rho for total number of passengers is now 0.06 (rather than 0.34), while the distance effect starts to emerge (rho is −0.15). In other words, even though the sociospatiality of a phenomenon like Sandy is expressed partly through a network of connections between territories, these connections are very much bounded by the locally-specific practices of place. So not only can we discern more complex sociospatial relations than just the immediate experience of a natural disaster through this data, but we can also understand how the spatially

[4] The red line through the scatterplot indicates a fit using a linear model. Confidence interval of the fit is indicated in light grey.

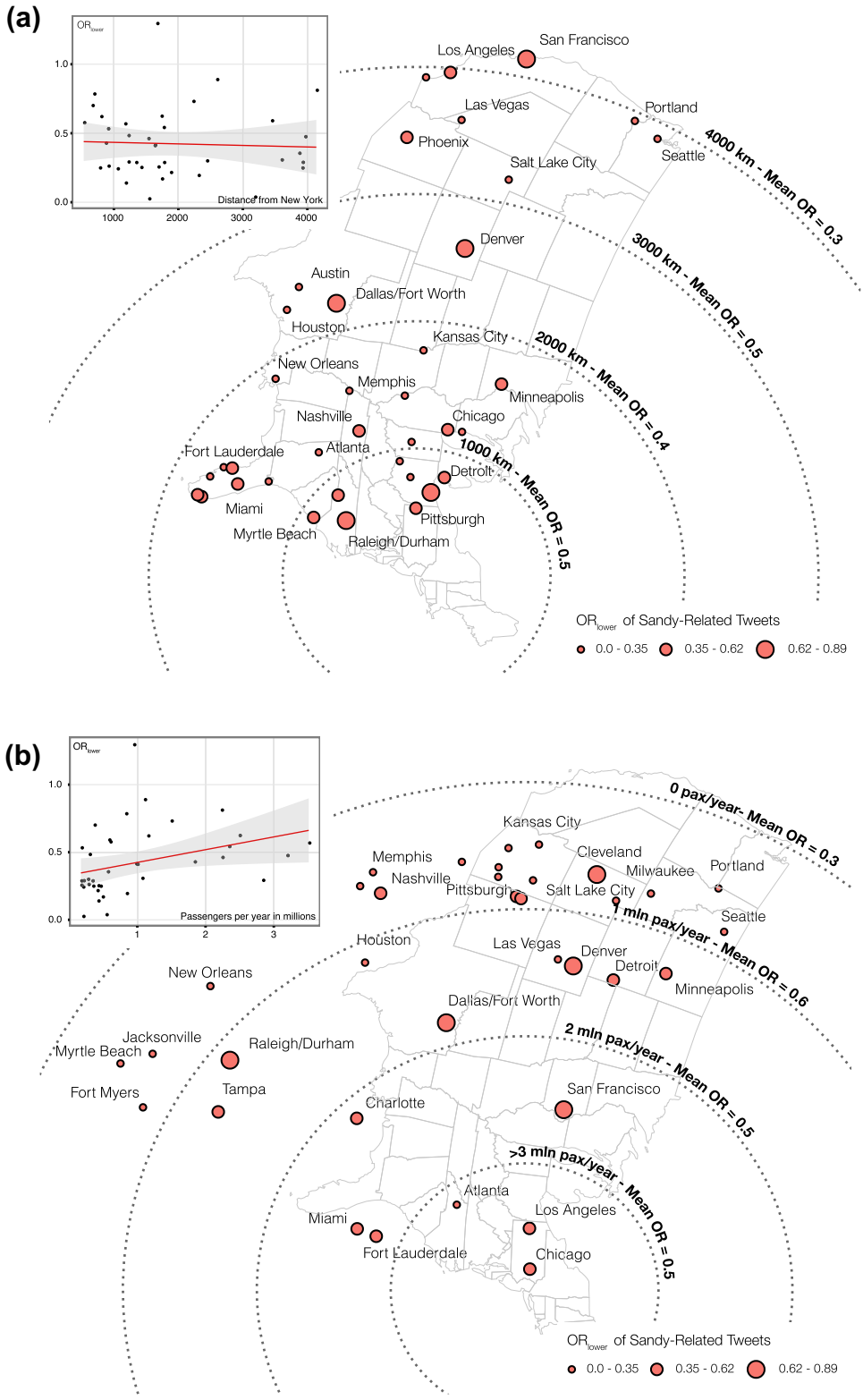

Fig. 6. Sandy's sociospatial networks.

distanciated networks have their own territorial groundings, just not only in those places one might expect. This once again highlights the complex ways in which the digital data shadows of a material event are manifest through the intertwining of different dimensions of social space.

As evidenced by each of these examples, Sandy's data shadows are not evenly distributed through the continental United States. They are instead quite intense in some locations, while hardly reaching other physically adjacent sites. Airline passenger movement partially explains how the data shadows of social processes are stretched over physical space and user-generated social media provides another indicator for better understanding the production of these relations. Disasters do not transpire in a single, unitary and bounded locale, but are embedded within complex and evolving

sociospatial relations that stretch unevenly across space. Some places are connected quite closely because of their political and economic interdependencies or dense social ties. Other places, while physically closer, which lack such substantive relational connections tend to have quite different experiences of such events.

5. Conclusion

The analysis of the data shadows of Hurricane Sandy presented here reveals relatively few surprises. Tweeting was largely concentrated in the areas hit hardest by the hurricane, with more distant areas having many fewer Sandy-related tweets. This analysis, however, has expanded via more holistic methodological and conceptual approaches, allowing us to demonstrate the shortcomings of simply plotting points on the earth's surface and assuming a one-to-one relationship between the location of tweets and the material events about which they are created. This kind of commonplace approach fails to acknowledge the unevenness of tweeting at different scales, it ignores the full range of knowledges represented in the content of tweets which themselves are locally specific, and it overlooks the spatially-distanciated, relational networks which complicate any assumptions of a uniform distance decay effect, among other things.

While there is undoubted potential in using social media in times of crisis, we worry that too much of the discourse and practice of crisis mapping, let alone other applications of this kind of data, relies on the relatively simple spatial ontologies and epistemologies that we have critiqued here. That is, seeing spatial concentrations of social media activity in disaster situations as being equivalent to areas in need of relief vastly oversimplifies the ways that social media is used in disaster situations, while also potentially reinforcing offline social inequalities by failing to provide relief to areas which may not be producing such content because of lack of access to the appropriate technologies or material conditions preventing the use of such tools (e.g., power outages). Geosocial media data can undoubtedly provide an important window into understanding disaster situations and formulating responses to them, but we would argue that any utilization of this data would be wise to account for the complexities that it embodies. While this need for problematization and caution might limit the usefulness of this kind of data in the immediate aftermath of disasters, disaster response is often a long and laborious process (as has been the case with Hurricane Sandy), so it is important to note that this data may well remain useful for analysis after the crisis event itself.

The promise of utilizing such big data sources for social scientific analysis is not solely in the size of the dataset, but the wealth of social processes that are encoded in such data. Thus, even though our case study of Hurricane Sandy does not present any radically new empirical insights into the geography of Twitter, or user-generated geographic information more broadly, we believe that this case study has allowed us to articulate three key conceptual and methodological points that should inform any similar analyses of geosocial media data in the future.

First, we have shown the utility of using small subsets of big data sources for social and spatial analysis. Starting with a large archive exceeding 3 billion geotagged tweets, we used only roughly 140,000 Sandy-related tweets for this case study. So even while this might have meant that there were just a few dozen data points in a given neighborhood in some cases, this amount of data is more than sufficient to gain statistically significant insights from our quantitative analysis, while also making qualitative analysis of these tweets more manageable. It is important to again emphasize that more data does not necessarily lead to more meaningful results, or a more accurate depiction of the world around us, something generally obscured by the contemporary fetish for 'bigness' in data.

Second, we have shown the importance of a mixed methods approach to understanding big data. A quantitative mapping of tweet density, however technically sophisticated, ultimately stops short of understanding the complex and polymorphous geographies of such data without also performing a qualitative analysis of the actual tweets and the context in which they are produced, or even employing a diversity of quantitative methods, such as social network analysis. Similarly, a qualitative analysis of such big data sources, even when narrowed down to just 140,000 data points, is impractical, if not impossible, without some a priori analysis and filtering based on quantitative methods. Such a mixed method approach not only avoids a kind of naïve empiricism with respect to big data that is currently prevalent, but also fundamentally points towards big data as embodying a variety of social and spatial relations which can begin to parse out through such analysis.

Finally, we have argued for the value of employing existing conceptual frameworks, such as Jessop et al.'s TPSN framework, to better understand the complexities of user-generated content and the sociospatial relations they embody. While most of the existing work on the geoweb has failed to explicitly theorize sociospatial relations, we have used the case of Hurricane Sandy and its data shadows to demonstrate the utility of the TPSN framework and its underlying analytical dimensions to produce much deeper understandings of space and spatiality as embodied in user-generated geographic information. We believe that these three contributions can help to provide a firmer foundation for future analyses of geosocial media data, highlighting the complex and variegated sociospatial relations represented in such data sources.

References

Amin, A., 2002. Spatialities of globalisation. Environ. Plan. A 34 (3), 385–399.

Anderson, C., 2008. The end of theory: the data deluge makes the scientific method obsolete. Wired Mag. 15 (7).

Boyd, D., Crawford, K., 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inform. Commun. Soc. 15 (5), 662–679.

Cairncross, F., 1997. The Death of Distance: How the Communications Revolution Will Change Our Lives. Harvard Business Press.

Carr, D.B., Olsen, A.R., White, D., 1992. Hexagon mosaic maps for display of univariate and bivariate geographical data. Cartogr. Geogr. Inform. Syst. 19 (4), 228–236.

Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W., Zook, M., 2013. Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. Cartogr. Geogr. Inform. Sci. 40 (2), 130–139.

Crawford, K., 2013. The hidden biases in big data. Harvard Business Review. 1 April. <http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html>.

Cresswell, T., 2004. Place: A Short Introduction. Blackwell.

Crutcher, M., Zook, M., 2009. Placemarks and waterlines: racialized cyberscapes in post-Katrina Google Earth. Geoforum 40 (4), 523–534.

Derudder, B., Witlox, F., Taylor, P., 2007. U.S. cities in the world city network: comparing their positions using global origins and destinations of airline passengers. Urban Geogr. 28 (1), 74–91.

Elwood, S., 2008. Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. GeoJournal 72 (3–4), 173–183.

Elwood, S., Goodchild, M.F., Sui, D.Z., 2012. Researching volunteered geographic information: spatial data, geographic research, and new social practice. Ann. Assoc. Am. Geogr. 102 (3), 571–590.

Floridi, L., 2012. Big data and their epistemological challenge. Philos. Technol. 25 (4), 435–437.

Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4), 211–221.

Goodchild, M.F., Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. Int. J. Digit. Earth 3 (3), 231–241.

Goodchild, M., Li, L., 2012. Assuring the quality of volunteered geographic information. Spat. Stat. 1 (1), 110–120.

Gormley, M., 2012. Cuomo: Sandy Cost NY, NYC $32b in Damage and Loss. The Associated Press, 26 November. <http://bigstory.ap.org/article/cuomo-bloomberg-brief-ny-delegation-storm-cost>.

Graham, M., 2012. Big data and the end of theory? The Guardian. 9 March. <http://www.guardian.co.uk/news/datablog/2012/mar/09/big-data-theory>.

Graham, M., 2013. The virtual dimension. In: Acuto, M., Steele, W. (Eds.), Global City Challenges: Debating a Concept, Improving the Practice. Palgrave, London, 117-139.

Graham, M., Hale, S.A., Gaffney, D., forthcoming. Where in the world are you? Geolocation and language identification in Twitter. Prof. Geogr.

Graham, M., Shelton, T., 2013. Geography and the future of big data; big data and the future of geography. Dialog. Hum. Geogr. 3 (3), 255–261.
Graham, M., Zook, M., 2011. Visualizing global cyberscapes: mapping user-generated placemarks. J. Urban Technol. 18 (1), 115–132.
Graham, M., Zook, M., 2013. Augmented realities and uneven geographies: exploring the geolinguistic contours of the web. Environ. Plan. A 45 (1), 77–99.
Graham, M., Zook, M., Boulton, A., 2013. Augmented reality in urban places: contested content and the duplicity of code. Trans. Inst. Brit. Geogr. 38 (3), 464–479.
Gregory, S., 2010. Cory Booker: the mayor of Twitter and Blizzard Superhero. TIME Magazine. 29 December. <http://www.time.com/time/nation/article/0,8599,2039945,00.html>.
Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environ. Plan. B: Plan. Des. 37 (4), 682–703.
Haklay, M., 2012. 'Nobody wants to do council estates': digital divide, spatial justice and outliers. Paper presented at the 108th Annual Meeting of the Association of American Geographers, February 25, 2012, New York, NY.
Hey, T., Trefethen, A., 2003. The data deluge: an e-science perspective. In: Berman, F., Fox, G., Hey, T. (Eds.), Grid Computing: Making the Global Infrastructure a Reality. John Wiley and Sons, pp. 1–17.
Hey, T., Tansley, S., Tolle, K., 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research.
Issenberg, S., 2012. The Victory Lab: The Secret Science of Winning Campaigns. Crown.
ITO World, 2010. ITO World at TED 2010 – Project Haiti. ITO World Blog. February 12. <http://itoworld.blogspot.com/2010/02/ito-world-at-ted-2010-project-haiti.html>.
Jessop, B., Brenner, N., Jones, M., 2008. Theorizing sociospatial relations. Environ. Plan. D: Soc. Space 26 (3), 389–401.
Kitchin, R., 2013. Big data and human geography: opportunities, challenges and risks. Dialog. Hum. Geogr. 3 (3), 262–267.
Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M., 2009. Computational social science. Science 323 (5915), 721–723.
Leetaru, K., 2011. Culturomics 2.0: forecasting large-scale human behavior using global news media tone in time and space. First Monday 16 (9).
Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E., 2013. Mapping the global Twitter heartbeat: the geography of Twitter. First Monday 18 (5).
Lehrer, J., 2010. A physicist turns the city into an equation. The New York Times. <http://www.nytimes.com/2010/12/19/magazine/19Urban_West-t.html>.
Leszczynski, A., 2012. Situating the geoweb in political economy. Prog. Hum. Geogr. 36 (1), 72–89.
Li, L., Goodchild, M.F., 2010. The role of social networks in emergency management: a research agenda. Int. J. Inform. Syst. Crisis Response Manage. 2 (4), 48–58.
Liu, S.B., Palen, L., 2010. The new cartographers: crisis map Mashups and the emergence of neogeographic practice. Cartogr. Geogr. Inform. Sci. 37 (1), 69–90.
Lohr, S., 2012. The age of big data. The New York Times. 11 February. <http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html>.
Manovich, L., 2011. Trending: the promises and the challenges of big social data. Debates in the Digital Humanities, 460–475.
Massey, D., 1991. A global sense of place. Marxism Today 35 (6), 24–29.
Mayer-Schonberger, V., Cukier, K., 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt.
Miller, H.J., 2010. The data avalanche is here. Shouldn't we be digging? J. Reg. Sci. 50 (1), 181–201.
Poorthuis, A., 2013. Getting Rid of Consumers of Furry Pornography, or How to Find Small Stories With Big Data. Working Paper.
Red Cross, n.d. The American Red Cross and Dell Launch First-Of-Its-Kind Social Media Digital Operations Center for Humanitarian Relief. <http://www.redcross.org/news/press-release/The-American-Red-Cross-and-Dell-Launch-First-Of-Its-Kind-Social-Media-Digital-Operations-Center-for-Humanitarian-Relief>.
Roche, S., Propeck-Zimmermann, E., Mericskay, B., 2013. GeoWeb and crisis management: issues and perspectives of volunteered geographic information. GeoJournal 78 (1), 21–40.
Ruppert, E., Law, J., Savage, M., 2013. Reassembling social science methods: the challenge of digital devices. Theory Cult. Soc. 30 (4), 22–46.
Scott, D.W., 1985. Averaged shifted histograms: effective nonparametric density estimators in several dimensions. Ann. Stat. 13 (3), 1024–1040.
Shelton, T., Zook, M., Graham, M., 2012. The technology of religion: mapping religious cyberscapes. Prof. Geogr. 64 (4), 602–617.
Smith, N., 2006. There's no such thing as a natural disaster. Understanding Katrina: Perspectives from the Social Sciences. Social Science Research Council. <http://understandingkatrina.ssrc.org>.
Stephens, M., Poorthuis, A., 2013. Follow Thy Neighbor: Connecting the Social and the Spatial on Twitter. Working Paper.
Takhteyev, Y., Gruzd, A., Wellman, B., 2012. Geography of Twitter networks. Soc. Networks 34 (1), 73–81.
Torrens, P., 2010. Geography and computational social science. GeoJournal 75 (2), 133–148.
Tsou, M., Leitner, M., 2013. Visualization of social media: seeing a mirage or a message? Cartogr. Geogr. Inform. Sci. 40 (2), 55–60.
Twitter, n.d. 2012 Year on Twitter: Global Town Square. <https://2012.twitter.com/en/global-town-square.html>.
Wald, D.J., Dewey, J.W., 2005. Did you feel it? Citizens contribute to earthquake science. United States Geological Survey Fact Sheet 2005-3016.
Wald, D.J., Quitoriano, V., Dengler, L.A., Dewey, J.W., 1999. Utilization of the Internet for rapid community intensity maps. Seismol. Res. Lett. 70 (6), 680–697.
Zook, M., Graham, M., Shelton, T., Gorman, S., 2010. Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. World Med. Health Policy 2 (2), 7–33.
Zook, M., Graham, M., Stephens, M., 2013. Data Shadows of an Underground Economy: Volunteered Geographic Information and the Economic Geographies of Marijuana. Unpublished Manuscript.