Beyond the Geotag? Deconstructing "Big Data" and Leveraging the Potential of the Geoweb

Jeremy W. Crampton* (University of Kentucky)
Mark Graham (Oxford Internet Institute)
Ate Poorthuis (University of Kentucky)
Taylor Shelton (Clark University)
Monica Stephens (Humboldt State University)
Matthew W. Wilson (University of Kentucky)
Matthew Zook (University of Kentucky)

*** Corresponding author (jcrampton@uky.edu)**

Beyond the Geotag? Deconstructing "Big Data" and Leveraging the Potential of the Geoweb

ABSTRACT

This paper presents an overview and initial results of geoweb analysis performed by our group as the foundation for a continued discussion of the potential impacts of "big" geodata for the practice of critical human geography. While Haklay's (2012a) observation that social media content is generated by a tiny fraction of participants ("outliers") is correct, we explore the methods and limits of going beyond geotagged datasets to overcome these issues. The results suggest a cautious approach towards the use of big geodata that is as mindful of its shortcomings as its potential.

The principal case study focuses on the widely reported riots following the University of Kentucky men's basketball team's 2012 championship, and its manifestation within the geoweb. Drawing upon a Twitter database we have been developing – which among other things has collected all geo-tagged tweets (about 5 million a day) since December 2011 – we analyze the geography of tweets that used a specific hashtag (#LexingtonPoliceScanner). The hashtag refers to the online feed of the Lexington Police Department (LPD) and by itself sheds light on spatially-determinable events as news of them diffuses over time and space.

We propose five extensions to the typical practice of mapping georeferenced data that we call going beyond the geotag: (1) examining social media that is not explicitly geographic; (2) spatialities beyond the "here and now"; (3) going beyond the proximate; (4) going beyond the human to data produced by bots and automated systems and (5) leveraging tweets against ancillary data (such as news reports and census data). These extensions create an "information-amplifier" effect that—at least partially—overcomes the partiality of geoweb social media.

We discuss our effort to develop a geoweb big data analytic engine which provides basic geovisualization functionality for a range of geo-tagged big data, enabling the combination of more conventional, "top-down" datasets or qualitative methods with user-generated geo-data.

Keywords: geoweb, #LexingtonPoliceScanner, big data, Twitter, geotag

Introduction

On April 2, 2012, following the victory of the University of Kentucky Wildcats men's basketball team in the NCAA championship game, a spontaneous celebration of fans spilled into the streets of Lexington, Kentucky, lasting well into the morning. That the street party became raucous was no surprise. Indeed, similar street parties had taken place over a decade earlier following similar championship victories, and an even more exuberant celebration had taken place just two days earlier following the team's victory over the rival Louisville Cardinals. But the celebrations on the night of April 2 were unique in that they were broadcast outside of Lexington, as a variety of users of the popular microblogging, social media platform Twitter took to the internet to relay the scenes of the riots as told by the Lexington Police Department's scanner (see Figure 1).

Figure 1, "Uh We have a partially nude male with a propane tank" Tweets

Using the #LexingtonPoliceScanner hashtag[1] (hereafter, "#LPS" for short, unless quoting from source) to organize the conversation, news of the riots spread quickly outside the bounds of the city in a way not previously experienced. Exactly 12,590 tweets were generated by 6,564 users using the #LPS hashtag, comprising a user-generated, geographically-referenced collective response to a local event, which we argue provides an excellent case study of both the potentials and pitfalls of using "big data" in geographic research.

Since Tim O'Reilly coined the term "Web 2.0" in 2005 to describe the growth in user-generated internet content (O'Reilly 2005), the emergence of Silicon Valley neologisms herald both massive investments in new technologies and the emergence of new functionalities and social practices built around such technologies, notably demonstrated by the emergence of social media platform of Twitter. Two of the most prominent of these recent trends include "location," or the introduction of geographically-aware computing into social networking applications, and "big data," signifying the collection and analysis of massive, cross-referenced databases about citizens and their activities.

While these concepts have driven billions of dollars of investment in start-ups and acquisitions (CB Insights 2012), they also feature centrally in the study of what has variously been termed the "geoweb," "neogeography" and "volunteered geographic information," concepts meant to signify a web 2.0-esque shift in the production of geographic information from official, expert sources to those of amateurs, volunteers or other previously marginalized producers of information (Crampton 2008). Indeed, many of the more prominent analyses of the geoweb have emphasized precisely the importance of location and big data, often employing massive

[1] Hashtags are text strings that are used to organize tweets from a diverse range of sources that all relate or speak to a central idea or theme.

databases of geotagged information, such as all of the content indexed by Google Maps, in order to study the geographies of this (often) user-generated internet content. We argue that such prominent analyses of the geoweb often suffer from two shortcomings. First, they fail to fully account for the limitations of "big data"-based analysis, and second, they remain too closely tied to the simplified spatial ontology of the geotag.

This paper is a call to think beyond such limited analyses of the geoweb and the now-popularized, simplistic visions of big data as an atheoretical solution to understanding the spatial dimensions of everyday life that are increasing well documented on the geoweb (see Anderson 2008 for the most notable example of this kind of thinking). To think *beyond the geoweb*, we suggest a reorientation of geoweb research in five key ways. First, we argue that the study of geoweb practices should go beyond simple visualizations of content using latitude/longitude coordinates. Second, we propose that geoweb research promote a perspective beyond the "here and now," an approach which attends to the significance of spatial relations as they evolve over time. Third, we point to the promise of analysis that is not limited to the explicitly-geographic dimensions of geoweb activity but includes a relational dimension, such as social network analysis. Fourth, we highlight the fact that geoweb content is not produced solely by human users, but is the product of a complex, more-than-human assemblage involving a diversity of actors, including automated content producers like Twitter spam robots. Finally, we highlight the importance of including non-user-generated data, such as governmental or proprietary corporate data sources, as a supplement in geoweb research.

The intent of this paper is not to call for an end to geoweb research, nor to offer definitive conclusions about what can be learned from such "big data" resources, but to offer an alternative

take on the possibilities and problems of the geoweb and to suggest fruitful avenues for future research. Put simply, our aim is provide a roadmap for analyzing the geoweb that goes beyond the geotag and its associated limitations.

Moving Beyond the Geotag

Since the widespread popularization of online mapping platforms and user-generated geographic information, often dated to the release of Google Earth in 2005, geographers have been at the forefront of studying the multiple geographies of the geoweb (Elwood 2010, 2011). Though some early research pointed to the possibility for new web-based forms of geographic information production to enable a democratization of GIS by way of the internet (Goodchild 2007), or the emergence of a new, more flexible ontology and epistemology for geographic information (Warf and Sui 2010), others saw more continuity than change. For instance, Elwood (2008) draws close parallels between the discourses around neogeography and those of the so-called GIS and Society debates of the 1990s (cf. Pickles 1995), in which a more socially-conscious approach to GIS was thought to at least ameliorate the usually massive power differentials between those with the ability to map and those who were being mapped. But such continuities ultimately point to the fact that the democratizing potential of the geoweb, like that of participatory GIS, is quite limited (Graham 2011; Haklay 2012b), and can often exacerbate existing inequalities. Building on this groundwork, research by Wilson (2011ab) has investigated the ways in which volunteered geographic information are implicated in processes of urban restructuring, as particular ways of seeing and coding the urban landscape digitally come to constrain the range of possible interventions in such a space. Leszczynski (2012) has similarly pointed to the possibility that the geoweb represents the neoliberalization of geographic information, as the production and distribution of such information is increasingly undertaken by

profit-driven corporations as opposed to an assumedly-benevolent state with a mandate for universal provision.

But perhaps more of interest for the purposes of this paper is a strand of research into the geographies of the geoweb that is primarily concerned with mapping the spatial contours of geo-coded internet information, including Google Maps placemarks (Graham and Zook 2011), Flickr photos (Hollenstein and Purves 2010; Wall and Kirdnark 2012), Wikipedia entries (Graham et al 2011) and, most recently, geocoded tweets. By aggregating and visualizing large databases of geoweb data, this research seeks to understand how these geolocated social media are connected to particular places and their cultural, economic, political and social histories. For instance, such research has shown how the distribution of dominant Christian denominations across the United States is reflected in online references within the Google Maps database (Zook and Graham 2010; Shelton et al 2013), as well as how the language in which geoweb content is produced can variously point to the centrality of place-based identities or the ways in which particular places are enrolled into global networks of tourism (Watkins 2012; Graham and Zook 2013). Such exercises have also been employed to more playfully map the diffusion of cultural memes across space, from zombies (Graham et al 2012) to the price of marijuana (Zook et al 2012). While these studies have provided an entry point for further work on the geographic dimensions of user-generated online content, and perhaps most importantly demonstrated the mutually constitutive nature of these spatially-referenced web 2.0 platforms and the offline social world, they have suffered from two primary faults.

First, the data used in these analyses are often quite limited in their explanatory value, no matter how “big” they might be. In an age in which massive datasets are often allowed to “speak

for themselves" (Gould 1981) with the hopes of providing some unforeseen insight into some aspect of human behavior, we argue that such studies, especially when drawing upon data collected by social media platforms, are naive in the way their insights are extrapolated to make sweeping statements about society as a whole (see boyd and Crawford 2012 for a discussion of these issues). Indeed, as Haklay (2012a) has argued previously, sources of big geosocial data are inherently biased towards "outliers". In other words, no matter how many geocoded tweets one is able to collect and analyze, they are a necessarily non-representative sample, as the number of geocoded tweets is but a small fraction of all tweets, and Twitter is used by only a small subset of all internet users, a group which itself represents only around one-third of global population (Graham 2012). As such, there is little that can be said definitively about society-at-large using only these kinds of user-generated data, as such data generally skews towards a more wealthy, more educated, more Western, more white and more male demographic. And while many of the aforementioned studies are quick to recognize and qualify their findings based on this limitation, it is especially important to maintain a skeptical position at a time in which the hype around "big data" is widespread.

Second, the aforementioned studies, while focusing explicitly on the geographic dimensions of user-generated content, employ a fairly simple spatial ontology, tied closely to the idea of "geotagging". By and large, these studies focus on the mapping of this user-generated content, relying on the attachment of "geotags", or associated latitude/longitude coordinates, in order to locate the placemarks, photos, wikis or tweets in geographic space. And while there is a certain importance to such an exercise—namely the verification of persistent digital divides in the production of internet content and the close connections between such online social activity and the offline world that is so often conceived of as being separate from it—we would argue

that such work displays an overreliance on geotags as a way of situating this data in geographic context, ignoring the multiplicity of ways that space is implicated in the creation of such data. For instance, a piece of information geotagged to a particular location may not necessarily have been produced in that location, be about that location, or exclude reference to any other geographic locality. Indeed, myriad examples suggest that geotagged content often exhibits a variety of spatial referents apart from this hidden latitude/longitude coordinates attached to it. Because of this, we argue that a more fully relational understanding of space (Massey 1991, 1993; Amin 2002) is necessary for understanding the production of geoweb content. Such a conceptual grounding allows us to emphasize that absolute location within the Cartesian plane of x/y coordinates belies the complexity of spatial relations between places as represented in the geoweb and the ways that the production of such geographically-referenced content is implicated in the production of space itself (Lefebvre 1991), and furthermore, that such content is co-constitutive with space ontogenetically (Kitchin and Dodge 2011).

But the promise of "big data" is not all for naught. Indeed, the massive amount of information captured, especially by social media platforms, offers an unprecedented opportunity to gather and analyze fine-grained data about social action. But the power of this data is especially acute when combined with other sources of information and a diversity of analytical techniques. In light of these limitations, we argue for a reconceptualization of the geoweb "beyond the geotag". That is, studies of the geoweb need to more completely take into account:

(1) social media data that is not explicitly geographic (e.g., tweets without geotags),

(2) spatialities beyond the "here and now" (e.g., scale-jumping, temporality)

(3) methodologies that are not focused on the explicitly proximate (e.g., relational connections, social network analysis)
(4) social media data that is not produced by humans (e.g. Twitter robots), and
(5) geographic data from non-user-generated sources (e.g., Census data)

These additions are not simply methodological refinements, but point to a distinctly different ontology and epistemology of the geoweb, one that more fully contextualizes this wealth of data within a broader range of socio-spatial practices than just static points on a map. By understanding the geoweb through a diversity of quantitative data sources and methodologies (e.g., mapping, spatial analysis, social network analysis), while also augmenting such analyses with in-depth qualitative analysis of users and places implicated in these data, we can understand the geoweb as something beyond a simple collection of latitude-longitude coordinates extraneously attached to other bits of information, and instead understand it as a socially-produced space that blurs the oft-reproduced binary of virtual and material spaces.

Similarly, we argue that critical geography more generally retains significant purchase for analyses of the geoweb, largely as a theoretical and methodological tool to push beyond an over-reliance on quantitative methodologies and data visualization while ignoring the context in which such data is produced. Rather than just asking where tweets are coming from, we can begin to contextualize these data and ask questions about what participation in geosocial networking means as a socio-spatial practice in the contemporary age. Such an approach acknowledges the limitations of letting the data speak for themselves, as is often advocated with so-called “big data”, but attempts to employ these data as one means among many for understanding the geoweb.

The Geographies of #LexingtonPoliceScanner

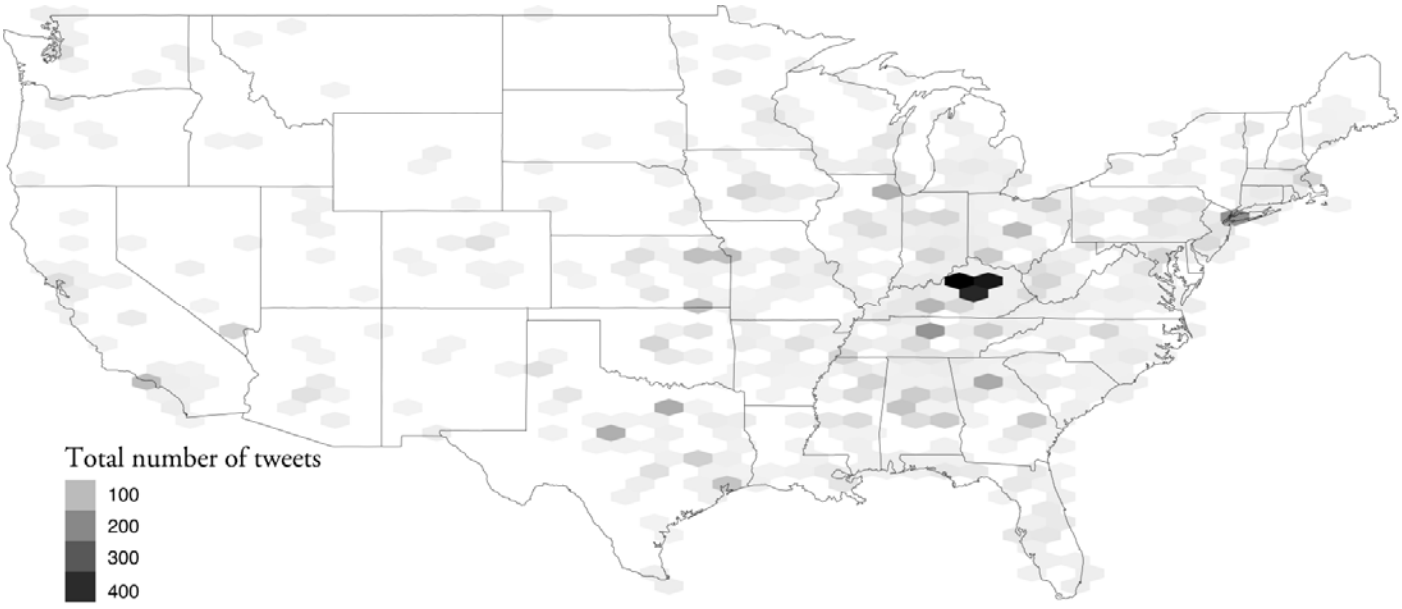
In order to demonstrate the utility of a program of geoweb research beyond the geotag, we offer an analysis of a short term, localized event in physical space that was well documented within geographic social media. The event -- an impromptu street party celebrating the victory of the University of Kentucky Wildcats men's basketball team in the NCAA championship game -- began in the late evening of April 2, 2012 and continued early into the next morning. At times, the celebration morphed into a riot as some fans set fire to couches and cars, threw bottles at police and fellow fans and otherwise engaged in a variety of criminal behavior.

A strong police presence and reaction prompted by the riots earlier on the weekend resulted in a sharp peak of chatter on the Lexington Police Department (hereafter, LPD) radio, which is accessible to the public either through both police scanners and online audio feeds. When listeners began to tweet events or quotations heard on the scanner using the #LPS hashtag, a short-lived Internet meme was born. For the purposes of this paper, we collected a comprehensive database of tweets using the #LPS hashtag via the Twitter streaming API. While the 12,590 tweets collected may not qualify under some definitions of “big data”, we see this dataset as providing us with a microcosm of the world of Twitter on which to base our analysis and critique. Through this analysis, we hope to illuminate the complexity, possibilities and shortcomings of “big data” research projects with an explicitly spatial focus by pointing to a variety of ways that geoweb research can move beyond the mere visualization of geotagged internet content.

1. Beyond the X/Y

When asking questions of large databases of geosocial media, the first and most obvious cut of the database is to map the basic spatial distribution of the phenomenon. Figure 2 below visualizes the geographic extent of the Twitter conversation referring to the #LPS hashtag. As one might expect, the level of interest in this event follows a classic distance decay function, with most discussion centered on Lexington, and dispersing outwards, especially towards larger cities nearby such as Nashville, Tennessee.

Figure 2, Distribution of All #LexingtonPoliceScanner Tweets[2]



This map, however, hides a number of issues that confound any one-dimensional mapping as any use of social media has a number of locations (e.g., sender, recipient, content, server, software, packet switching paths, etc.) that might be relevant in a given analysis (Zook and Dodge 2009). Moreover, data on some of these locational characteristics are relatively easy

[2] Please note that these maps are currently unprojected and in low resolution for the purposes of this review. We can provide corrected, high resolution versions upon acceptance.

to obtain, while others are nearly impossible to collect in a systematic manner, only further compounding the problem. In the case of data from the Twitter API, we are able to access either the location of the user or the location of the tweet. The former is based on a user-specified location in one's profile and is unverified. Users can provide a variety of types of locations, ranging from latitude/longitude coordinates in decimal degrees to city or country names to fictional locations such as "Middle Earth". This fuzziness makes geocoding user location problematic, although approximately 60 percent of all tweets can ultimately be associated with a physical location with some degree of confidence (Hale et al 2012). A major disadvantage of this approach is that it divorces the (usual) location of the user from the location in which a tweet is created. For example, someone might list their location as Goshen, Indiana but tweet from Hesston, Kansas. While such disjunctures are interesting in that they represent an alternative relationship between geotagged internet content and social practice through the association of multiple locations, they nonetheless represent a limitation to conventional forms of locating user-generated content in geographic space.

An alternative method for locating geosocial media is the information associated with the actual tweet itself. Derived from GPS coordinates in a mobile device or triangulation from cell or Wi-Fi signals, a geocoded tweet provides the site where the act of tweeting occurs. While this provides a great deal more confidence in location, it is not without its own problems. Most significant is the fact that users have to opt-in in order to provide this form of location information, and as a result only about one percent of all tweets are geocoded. Moreover, the scale at which this geocoding is accurate varies, which has further implications for the scale at which this data is analyzed.

In the case of the #LPS tweets, we found that 34 percent of the 12,590 tweets in the original database were geocodable by user-defined location information in profiles while only 0.2 percent were geocodable by tweet location. While this is an obvious limitation to understanding the geography of #LPS tweets, it does not render this data irrelevant. It does, however, necessitate caution in choosing how to go about analyzing the data, and points to the importance of expanding the analysis beyond the simple mapping of points in space. For example, the user-supplied location information we use in the analysis provided by Figure 2 is but the first slice that can be taken from this big data, dataset.

2. Beyond the "Here and Now"

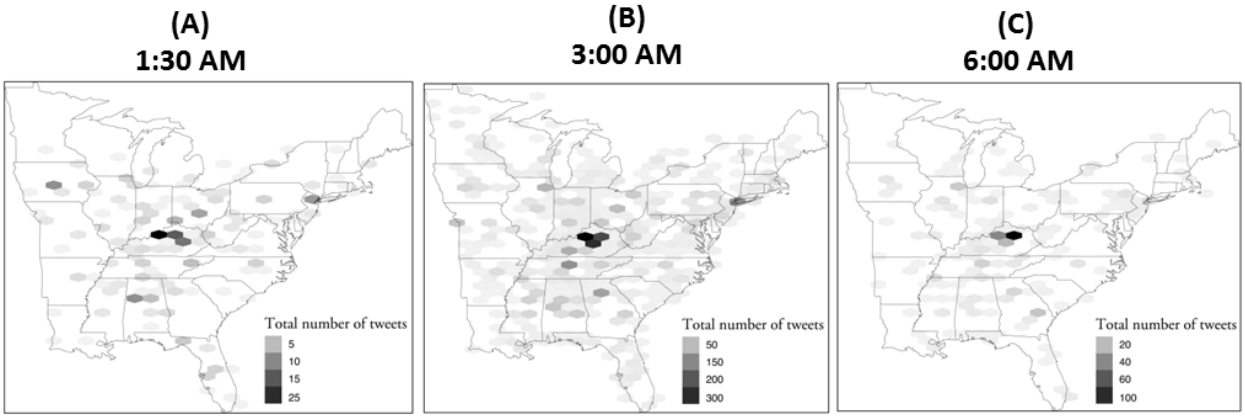
Like many other cities around the United States, the Lexington police scanner is streamed online, but does not have a large regular audience. However, on Monday evening, April 2nd, at 11:50:49pm, a Twitter user transcribed an audio clip of the police scanner noting that shots had been fired, adding the hashtag #LPS. This act, and the many other #LPS tweets and retweets that followed, effectively broadcast the Lexington police scanner beyond the local, the "here and now", diffusing the news across the country and globe, jumping scales from a local occurrence to a worldwide phenomenon and briefly becoming a globally trending topic on Twitter.

Figure 3, "Shots fired" Tweet

This occurrence highlights the need to look beyond static representations of geoweb data and consider the space-times of geodata diffusion. Going beyond the static visualization of all

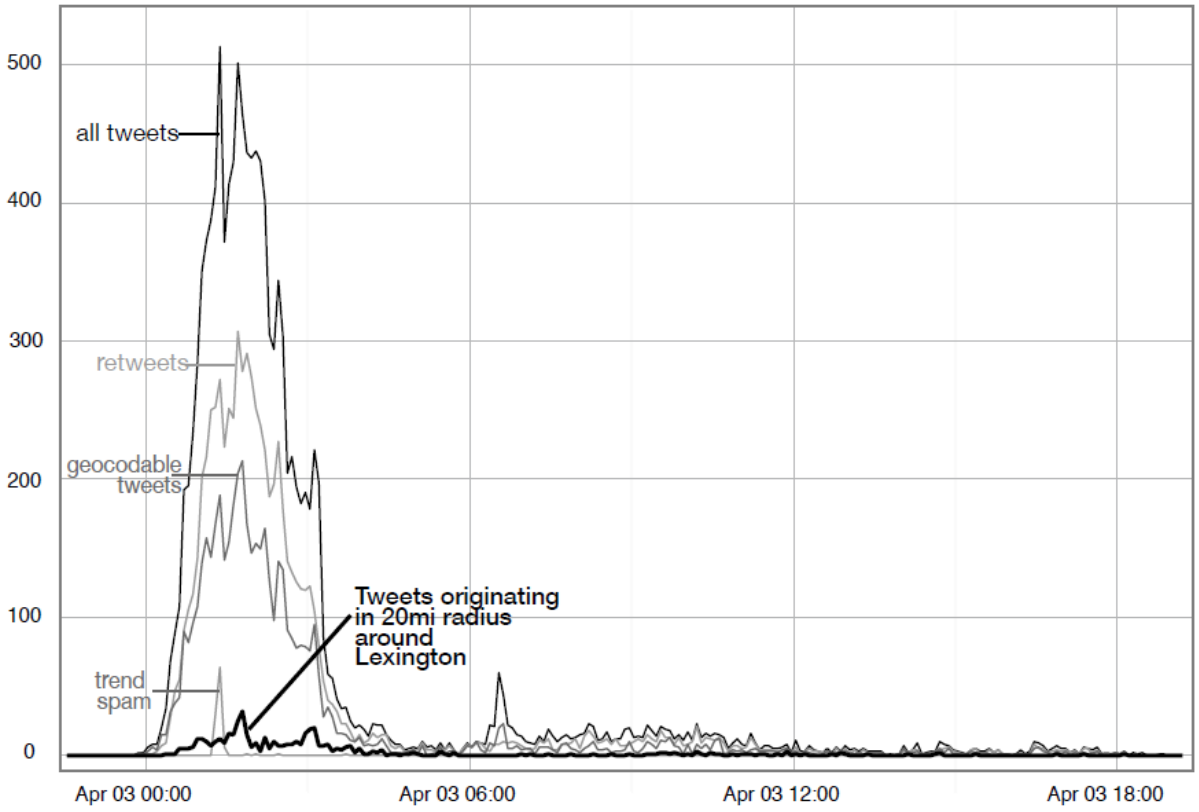
#LPS tweets in Figure 2, Figure 4 illustrate how the spatial patterns in tweeting differ from over time. The original attention for the police scanner audio stream emerged from the region around Lexington (Figure 4a), though notably not from within Lexington itself. The event was subsequently picked up throughout the US (and in Lexington itself) creating a Twitter "trend". After approximately two hours (Figure 4b), national interest in the trend peaked and began to decrease, leaving only the region around Lexington to tweet about the event (Figure 4c).

Figure 4, The Geographic Distribution of Tweets



In addition to changing spatial extent of #LPS tweeting, Figure 5 highlights how the frequency of these tweets evolved over time. The black line indicates all #LPS tweets. Looking at the temporal dimension reveals that the event very quickly (within one hour) became a trending topic, but that the actual trend was relatively short-lived. Only three hours after the first tweet was sent, attention died down, only peaking again for a brief time around 6:30 am after the social media news blog Mashable reported about the events of the previous night. Moreover, when one begins to disaggregate the frequency of tweets by type additional patterns emerge.

Figure 5, The Frequency of Tweets Over Time



For example, the line reflecting the number of retweets[3] (forwarding a received tweet rather than creating new content) with sixty percent of the dataset were literal copies of another tweet, while many more were slightly modified tweets without giving "official" attribution. Especially in the later stages of the night, almost the entire corpus of messages are retweets of tweets sent earlier in the evening suggesting that this trending event within social media was rather thin in original content.

It is also worth noting how much information would be lost were we to limit our analysis to the conventional X/Y coordinates. Nevertheless, geocoded information can provide useful

[3] In this context retweets are only those tweets that start with "RT."

insight as the line representing the frequency of tweets within a 20 mile radius around Lexington exhibits a different pattern than the general trend which corresponds to the findings of Figure 4. Attention to the police scanner peaked much later in Lexington than elsewhere, but the relative share of tweets generated within Lexington increased towards the end of the night. A final temporal anomaly is the example of trend spam which shows how specific actants operating at cross purposes to the original trends can only be fully appreciated by looking beyond the here and now.

3. Beyond the Proximate

An additional dimension to the geographies of #LPS, beyond the changing pattern of tweets over space and time, is the relational connections that collectively constitute the dataset. That is, the social networks through which ideas are developed and discussion is propagated, which represent a key means by which knowledge is transferred in a networked society. Building upon the work of sociologists, social network analysis allows one to look at the level and frequency of connections between individuals. This provides insight distinct from simply examining how spatially proximate things might be. While we are in no way arguing that physical distance is irrelevant, it is important to acknowledge forms of relational, cultural or linguistic distance that might not be so easily measured in kilometers or miles. Two people might be quite far apart from each other in physical distance, but be able to maintain an extremely close friendship utilizing social media platforms such as Twitter.

Figure 6 illustrates two retweet networks. The blue network connects a tweet made by one of the earliest listeners to the police scanner stream. Just past midnight, the Twitter user @TKoppe22 tweeted “Uh We have a partially nude male with a propane tank

#LexingtonPoliceScanner" from Knoxville, TN. In the next four hours, more than 200 people retweeted that message across the United States (see Figure 1 for the original tweet and some of the subsequent retweets). Figure 6 connects the location of every retweet with the location of the original tweet by @TKoppe22. Although there are strong links with both Louisville and Lexington, retweets are also spread across the entire eastern United States. The red network visualizes a similar retweet network but for the tweet "#LexingtonPoliceScanner is trending" first sent by DavidWood90 located in Chapel Hill, NC. In sharp contrast to the retweet network of the partially nude male, this tweet is picked up specifically in and around Lexington, perhaps pointing to a feeling of pride that what was expected to be a locally-confined event had become a global trend.

Figure 6, Retweet Networks Overlaid Onto Physical Space

Blue= Retweeting of "We have a partially nude male with a propane tank";
Red = Retweeting of “#LexingtonPoliceScanner is trending worldwide.”

This simple example illustrates the complicated relationship between physical and social distance. Moreover it shows that while physical distance remains important, social networks differentially stretch across physical space, connecting and reconnecting locales based on the message and the motivations of the social interaction.

4. Beyond the Human

We argue that another complementary way of moving beyond the geotag is to recognize the role of content and information that is *not* generated by the users typically under examination. A small subset of activity within the #LPS dataset, and Twitter data as a whole, is produced by “Twitter robots” (or “bots”), Twitter accounts that watch for trending hashtags and then tweet using that hashtag in order to capture attention and direct it toward specific advertised products. As shown in Figure 5, the “trend spam” line in the graph indicates a bot account that issues tweets in order to sell iPhones. Just as original tweets (black) and re-tweets (grey) peak in the early hours of Tuesday morning, bots measure these trends and respond. For instance, one bot tweeted at 1:18am: “Yay, got my iPhone4 delivered and its free! Cant believe it, see if you can get one 2 http://t.co/aOriI16m #LexingtonPoliceScanner,” while other bots attempted to sell sandals: “#LexingtonPoliceScanner #TodayDeals Sandal Sale! Save $30 on $100 Sandal orders plus free shipping from ShoeMall http://t.co/lTm3PIxO.”

This content is not user-generated (although some bot-tweets might actually be viruses that infect accounts of real users) but automated content produced by more-than-human

accounts. Lines of code automatically produce content tagged with #LPS (or any other hashtag) in order to capitalize on the attention paid to trending topics. Having little to nothing to do with the events taking place in Lexington, this content highlights the myriad ways in which software automatically produces content, and the ways in which the attention to the hashtag produces a marketable terrain on which some individuals can capitalize.

The code platform, produced by Twitter, operates as an actant, enabling manipulation within their broadcasts by gathering trending data and highlighting these trends at various geographic scales. Indeed, the automated attention of a bot is only triggered through a rise in trends at particular scales (and locations?), which is activated through Twitter's platform of tweets and retweets, producing new activity in the network even despite the presence of new tweets (see the gap in the black and grey line in the above graph). To move beyond the human in the study of the geoweb is to recognize and inquire into the variegated assemblages that went into producing the #LPS phenomena, including the various more-than-human elements that participated.

5. Beyond the User-Generated

Following the need to understand how non-human sources contribute to the otherwise user-generated content found on Twitter, we similarly argue for a need to look beyond sources of typically user-generated content in analyses of the geoweb. Because of the significant limitations to who and what Twitter data, or any geosocial media data sources, represent (Haklay 2012a), it is important to look beyond and leverage these sources, even when the constitute the primary area of focus for a given study.

In contrast to work typically carried out by analysts of "big data" who emphasize the massive quantity and unceasing flow of data (the "firehose" of data) and solutions for processing large data sets, we emphasize that the informational richness of these data is often lacking. In other words, while it is often high in quantity, it is not necessarily equally high in quality. Naturally, big data will often yield insights in and of itself, but we argue that by leveraging the available user-generated data (in this case the #LPS tweets) with other datasets, and by marrying and tracing interactions between user-generated data and events outside the users' knowledge or control, that an additional richness is provided to an analysis otherwise impossible by limiting oneself to single data sources. Indeed, we would argue that it is absolutely *necessary* to leverage user-generated data with ancillary datasets if one is to maximize the utility of the data.

As we have seen, following the very first tweet with the #LPS hashtag at 11:50:49PM, more than 12,000 additional tweets followed overnight. Significant proportions of these tweets did not add new commentary, but instead were retweets of earlier comments. We do know, however, that during this time, the web-based broadcast of the LPD was being accessed by listeners. In reading the content of the tweets that quoted the police scanner, it is evident that they tended to focus on the sensationalist and dramatic statements (the semi-nude man, shots fired, or the attractive voice of the female dispatcher, for example). News media too described scenes verging on the riotous with "dozens" of people arrested and a man wounded by "gunfire."

These events, however, can hardly be seen as representative of the entirety of the evening. It is here necessary to go beyond the abundance of user-generated tweets available and "ground truth" them by examining external, supplementary data sources, for example, LPD crime data on the number and location of arrests made that night, in order to build a more

comprehensive picture of the evening's events. For example, LPD records of the evening in question yields only four crime incidents at State Street (the center of the event) and zero on campus nearby. Even expanding the time range to include the day before yields fewer than 20 incidents, some clearly occurring outside the temporal extent of the evening's events. Clearly a globally (albeit briefly) trending Twitter topic can often imprecisely correlate to more locally grounded sources.

Another example is the small spike in tweets with the #LPS hashtag well after the events of the evening were over, at approximately 6:30AM the following morning. We have traced this to news media reporting on the #LPS hashtag, and specifically to the first media report, which came from the Mashable.com site (Laird 2012). This in turn was then retweeted, causing an observable spike in the data. This demonstrates an interesting self-sustaining interaction: news reports pick up on the tweets, and then the tweets pick up on the news reports. (We found other tweets also performing a kind of meta-comment, for example noting that #LPS is trending, with people then retweeting those comments.)

In short, studies utilizing big geoweb data would be well served by comparing and combining it with other data sources such as police reports (in the case of #LPS) or perhaps standard census data. There is, however, another form of ancillary data relevant to our case study, namely user collected imagery from the physical site of event. Perhaps most immediately relevant for estimating crowd size, this technique can provide extremely useful material for counter narratives. Although we can consult news reports to estimate the size of crowds, for example during Occupy Wall Street protests, these are often inaccurate. Instead, participants can fly a drone over the crowd to collect imagery from which the crowd's size may be estimated.

Then, given the crowd's size, it is possible to estimate the incidence of arrests, which as we have seen from official crime reports would be rather low, in contrast to the sensationalist content reported on Twitter and the news media.[4]

Conclusion

The goal of this paper has been to set forth a series of future directions for geoweb research, focusing primarily on moving beyond the simple mapping and analysis of user-generated online content tagged to particular points on the earth's surface. Instead, we have suggested that a closer attention to the diversity of social and spatial processes, such as social networks and multi-scalar events, at work in the production, dissemination and consumption of geoweb content provides a much fuller analysis of this increasingly popular phenomenon. That being said, even the preliminary analysis presented here to demonstrate the utility of such approaches is not comprehensive, nor definitive. Indeed, a variety of further avenues of research are equally promising, including micro-ethnographies of Twitter users across their lifespan, meant to produce genealogies of content production over time in order to contextualize involvement in particular events such as #LPS. Such analyses also point to the possibilities of greater integration between GIScience and critical human geography, as both have much to contribute to understanding the multiple dimensions of contemporary phenomena like the geoweb.

We also urge caution regarding the surveillant potential of this research. Although we did not do so here, it is today technically possible to not only follow a single tweet and its retweets

[4] Although we did not have the resources in place for the events of April 2012, our group recently performed a drone flight over another UK basketball event which involves fans camping out in tents on the campus in order to get tickets for an important game. From the imagery we can estimate the crowd size.

through the network (see Figure 6) but to trace who retweets which tweet. When someone retweets, did they retweet the original tweeter or another retweet? Our data only indicate that a tweet has been retweeted; the complex pattern of relationships among tweets is not derivable from this particular dataset. But it is being done. Why is this important? Using this more powerful analysis it is possible to disambiguate and identify opinion leaders and key influencers through behavioral analysis. In a crowded environment, who is driving opinion, who is spreading it, and who is acting more passively?

These questions become more germane given the interest in open source social media by western intelligence agencies in order to predict events and identify terrorists. The United States intelligence community has made several substantial steps in this direction in recent years. For example, the Office of the Director of National Intelligence, which oversees the US's 16 intelligence agencies (which collectively spend about $75 billion per year), maintains an Open Source Center (OSC) which analyzes social media forums such as Facebook and Twitter. According to Patrick O'Neil, Director of Analytic Development at OSC, open source intelligence (OSINT) is the "hot new field" in intelligence gathering. Letitia Long, who directs the nation's top geographical intelligence agency (which provided imagery, maps and even a model mockup of bin Laden's compound), has stated publicly that she sees a shift from traditional (clandestine) intelligence gathering to not only OSINT, but social media intelligence. Her agency, the National Geospatial-Intelligence Agency (NGA) recently released their Strategic Plan, Objective 2 of which is to "leverage and exploit…non-traditional (e.g., human geography and social media) geospatial sources" (NGA, 2012: 9). Similar efforts are underway in other countries; for example in the UK MI5 and GCHQ (the UK government communications headquarters where

communication signals are intercepted and analyzed) recently issued a call for research in identifying key opinion leaders and terrorist threats in "crowded environments".[5]

Furthermore, the reference to "human geography" by the NGA is part of an extremely significant effort by the intelligence community to supplement traditional information sources (such as surveillance satellites) with "socio-cultural analysis" (SCA) and the "human terrain." Here, geographical concepts and basic theories are being enrolled into military training both prior to and during deployment, so that, for example, open source social media content can be sifted to identify threats in hostile environments. The surveillant and monitoring aspects of this are undeniable implications of the research we report here.

Our analysis is forward-looking and programmatic. It is not meant to be an attack on geoweb analysis. Rather, we have been inspired by Haklay's well-reasoned comments concerning the tendency to focus only on outliers (Haklay 2012a,b) despite their lack of representativeness. We seek here to maximize the utility of geoweb studies, enrich the data-sets and analysis that can be produced, and to overcome limits of the data and finally to more fully let the data "speak for themselves" (Gould 1981).

[5] See http://www.science.mod.uk/events/event_detail.aspx?eventid=177

Citations

Amin, Ash. 2002. “Spatialities of globalisation.” *Environment and Planning A* 34(3): 385 – 399.

Anderson, Chris. 2008. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” *Wired Magazine* 16(7).

boyd, danah, and Kate Crawford. 2012. “Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon.” *Information, Communication & Society* 15(5): 662–679.

CB Insights. 2012. “Big Data Companies Pull $1.1B from Venture Capitalists Since Q2 2011.” *CB Insights Research.* Available from: http://www.cbinsights.com/blog/venture-capital/big-data-companies-venture-capital-fundinc (Last accessed 9/27/2012).

Crampton, Jeremy. 2008. “Cartography: maps 2.0.” *Progress in Human Geography* 33(1): 91–100.

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

Elwood, Sarah. 2010. “Geographic information science: Emerging research on the societal implications of the geospatial web.” *Progress in Human Geography* 34(3): 349–357.

Elwood, Sarah. 2011. “Geographic information science: Visualization, visual methods, and the geoweb.” *Progress in Human Geography* 35(3): 401–408.

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

Gould, Peter. 1981. “Letting the data speak for themselves.” *Annals of the Association of American Geographers* 71(2): 166-176.

Graham, Mark. 2011. Wiki Space: Palimpsests and the Politics of Exclusion. In *Critical Point of View: A Wikipedia Reader*, eds. G. Lovink and N. Tkacz. Institute of Network Cultures. pp. 269-282.

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

Graham, Mark, Scott A. Hale, and Monica Stephens. 2011. *Geographies of the World's Knowledge*. Convoco! Edition.

Graham, Mark, and Matthew Zook. 2011. "Visualizing Global Cyberscapes: Mapping User-Generated Placemarks." *Journal of Urban Technology* 18(1): 115–132.

Graham, Mark, and Matthew Zook. 2013. "Augmented Realities and Uneven Geographies: Exploring the Geo-linguistic contours of the web". *Environment and Planning A.*

Graham, Mark, Taylor Shelton, and Matthew Zook. 2012. "Mapping Zombies". In *Zombies in the Academy*, forthcoming.

Hale, Scott A., Devin Gaffney and Mark Graham. 2012. "Where in the world are you? Geolocation and language identification in Twitter."

Hollenstein, Livia, and Ross S. Purves. 2010. "Exploring place through user-generated content: Using Flickr tags to describe city cores." *Journal of Spatial Information Science* 1(1): 21–48.

Kitchin, Rob, and Martin Dodge. 2011. *Code/Space: Software and Everyday Life*. The MIT Press.

Laird, Sam. 2012. "#LexingtonPoliceScanner: Twitter Listens, Reacts to Kentucky Riots". *Mashable.com*, 3 April 2012. Available from: http://mashable.com/2012/04/03/lexingtonpolicescanner-twitter-listens-reacts-to-kentucky-riots-pics/ (Last accessed 9/27/2012).

Lefèbvre, Henri. 1991. *The Production of Space*. Blackwell.

Leszczynski, Agnieszka. 2012. "Situating the geoweb in political economy." *Progress in Human Geography* 36(1): 72–89.

Haklay, Muki. 2012a. "'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. New York, NY. February 25, 2012.

Haklay, Muki. 2012b. "Neogeography and the delusion of democratisation". *Environment and Planning A*, forthcoming.

Massey, Doreen. 1991. "Global sense of place". *Marxism Today*. June 1991.

Massey, Doreen. 1993. "Power-geometry and a progressive sense of place". In *Mapping the futures: Local cultures, global change*, eds. J. Bird, B. Curtis, T. Putnam, G. Robertson and L. Tickner. Routledge. pp. 59-69.

National Geospatial-Intelligence Agency, 2012. *NGA Strategy 2013-2017*. Springfield, VA: NGA.

O'Reilly, Tim. 2005. "What Is Web 2.0." *O'Reilly Media*. http://oreilly.com/pub/a/web2/archive/what-is-web-20.html.

Pickles, John. 1995. *Ground Truth: The Social Implications of Geographic Information Systems*. Guilford Press.

Shelton, Taylor, Matthew Zook, and Mark Graham. 2013. "The Technology of Religion: Mapping Religious Cyberscapes." *The Professional Geographer* 65(1): 1–16.

Wall, Melissa, and Treepon Kirdnark. 2012. "Online Maps and Minorities: Geotagging Thailand's Muslims." *New Media & Society* 14(4): 701–716.

Warf, Barney, and Daniel Sui. 2010. "From GIS to neogeography: ontological implications and theories of truth." *Annals of GIS* 16(4): 197–209.

Watkins, Derek. 2012. "Digital Facets of Place: Flickr's Mappings of the U.S.-Mexico Borderlands". Unpublished M.A. Thesis, University of Oregon Department of Geography.

Wilson, Matthew W. 2011a. "'Training the eye': formation of the geocoding subject." *Social & Cultural Geography* 12(4): 357–376.

Wilson, Matthew W. 2011b. "Data matter(s): legitimacy, coding, and qualifications-of-life." *Environment and Planning D: Society and Space* 29(5): 857–872.

Zook, Matthew, and Martin Dodge. 2009. "Mapping, Cyberspace." In *International Encyclopedia of Human Geography (Volume VI),* eds. R. Kitchin and N.Thrift. Elsevier. pp. 356-367.

Zook, Matthew, and Mark Graham. 2010. "Featured graphic: The virtual 'bible belt'." *Environment and Planning A* 42(4): 763–764.

Zook, Matthew, Mark Graham, and Monica Stephens. 2012. "Data Shadows of an Underground Economy: Volunteered Geographic Information and the Economic Geographies of Marijuana". Unpublished manuscript.