Cartography and Geographic Information Science, 2013
Vol. 40, No. 2, 130–139, http://dx.doi.org/10.1080/15230406.2013.777137

Taylor & Francis
Taylor & Francis Group

Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb

Jeremy W. Crampton[a]*, Mark Graham[b], Ate Poorthuis[a], Taylor Shelton[c], Monica Stephens[d], Matthew W. Wilson[a] and Matthew Zook[a]

[a]*Department of Geography, University of Kentucky, Lexington, KY, USA;* [b]*Oxford Internet Institute, University of Oxford, Oxford, UK;* [c]*Graduate School of Geography, Clark University, Worcester, MA, USA;* [d]*Department of Geography, Humboldt State University, Arcata, CA, USA*

(*Received 1 October 2012; accepted 29 January 2013*)

This article presents an overview and initial results of a geoweb analysis designed to provide the foundation for a continued discussion of the potential impacts of 'big data' for the practice of critical human geography. While Haklay's (2012) observation that social media content is generated by a small number of 'outliers' is correct, we explore alternative methods and conceptual frameworks that might allow for one to overcome the limitations of previous analyses of user-generated geographic information. Though more illustrative than explanatory, the results of our analysis suggest a cautious approach toward the use of the geoweb and big data that are as mindful of their shortcomings as their potential.

More specifically, we propose five extensions to the typical practice of mapping georeferenced data that we call going 'beyond the geotag': (1) going beyond social media that is explicitly geographic; (2) going beyond spatialities of the 'here and now'; (3) going beyond the proximate; (4) going beyond the human to data produced by bots and automated systems, and (5) going beyond the geoweb itself, by leveraging these sources against ancillary data, such as news reports and census data. We see these extensions of existing methodologies as providing the potential for overcoming existing limitations on the analysis of the geoweb.

The principal case study focuses on the widely reported riots following the University of Kentucky men's basketball team's victory in the 2012 NCAA championship and its manifestation within the geoweb. Drawing upon a database of archived Twitter activity – including all geotagged tweets since December 2011–we analyze the geography of tweets that used a specific hashtag (#LexingtonPoliceScanner) in order to demonstrate the potential application of our methodological and conceptual program. By tracking the social, spatial, and temporal diffusion of this hashtag, we show how large databases of such spatially referenced internet content can be used in a more systematic way for critical social and spatial analysis.

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

Downloaded by [Jeremy Crampton] at 05:18 19 April 2013

Introduction

On 2 April 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 2 April 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).

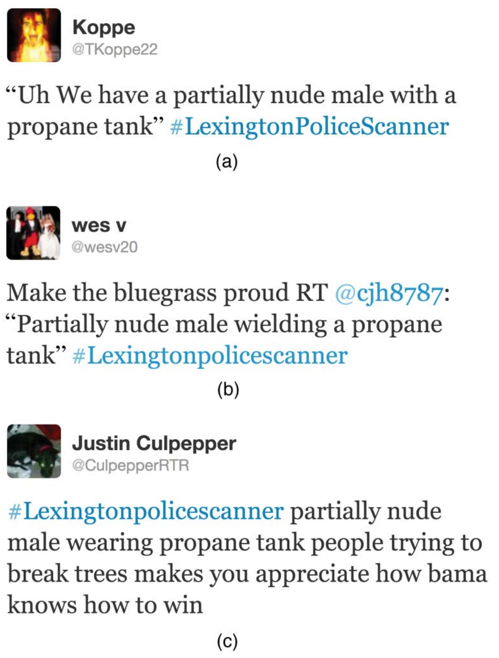
Using the #LexingtonPoliceScanner hashtag[1] (hereafter, '#LPS' for short, unless quoting from source) to organize the conversation, news of the riots spread quickly outside the bounds of the city in a way not previously experienced. Exactly 12,590 tweets were generated by 6564 users using the #LPS hashtag, 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 so-called 'big data' in geographic research. While this data set may not itself qualify as 'big' under many definitions – indeed there remains a significant contention around what exactly qualifies as big data (for instance, see Laney [2001], with regard to volume, velocity, and variety) – it is used to demonstrate the persistent importance of smaller subsets of data, even when larger data sets may be available, while also highlighting the utility of our methodological and conceptual program.[2]

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

*Corresponding author. Email: jcrampton@uky.edu

© 2013 Cartography and Geographic Information Society

Downloaded by [Jeremy Crampton] at 05:18 19 April 2013



Koppe
@TKoppe22

"Uh We have a partially nude male with a propane tank" #LexingtonPoliceScanner

(a)

wes v
@wesv20

Make the bluegrass proud RT @cjh8787: "Partially nude male wielding a propane tank" #Lexingtonpolicescanner

(b)

Justin Culpepper
@CulpepperRTR

#Lexingtonpolicescanner partially nude male wearing propane tank people trying to break trees makes you appreciate how bama knows how to win

(c)

Figure 1. 'Uh We have a partially nude male with a propane tank' tweets.

neologisms has heralded 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 (Crampton 2008), and the aforementioned 'big data', signifying the collection and analysis of massive, cross-referenced databases about citizens and their activities. But as these ideas have been taken up in more academic analyses, we argue that they tend to suffer from two primary shortcomings; first, they fail to fully account for the limitations of a big data-based analysis, and second, they remain too closely tied to the simplified spatial ontology of the geotag.

This article 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 increasingly well documented on the geoweb (see Anderson 2008 for the most notable example of this kind of thinking). To think beyond the geotag 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 promotes 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 article is not to call for an end to geoweb research or the use of big data, nor to offer definitive conclusions about what can be learned from such resources, but to offer a programmatic, alternative take on the possibilities and problems of the geoweb and big data and to suggest fruitful avenues for future research. Put simply, our aim is to 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. Here, scholars have examined the limits to the democratizing potential of the geoweb (Graham 2011; Haklay 2013), as well as the ways in which new iterations of geographic data are enrolled in broader political-economic processes (Wilson 2011a, 2011b; Leszczynski 2012).

The geographies of the geoweb have been further examined through mapping the spatial contours of geocoded internet information, including Google Maps placemarks (Graham and Zook 2011), Flickr photos (Hollenstein and Purves 2010; Wall and Kirdnark 2012), Wikipedia entries (Graham, Hale, and Stephens 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

Downloaded by [Jeremy Crampton] at 05:18 19 April 2013

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, Zook, and Graham 2012), 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, Shelton, and Zook 2013) to the price of marijuana (Zook, Graham, and Stephens 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 data sets are often thought to 'speak for themselves' (Gould 1981) without intervention, we emphasize the need to choose the appropriate data, to tease out inherent patterns and trends through data mining, and to share and explain those patterns in intriguing ways with the hope of providing some unforeseen insight into some aspect of human behavior. That is, big data will not replace thinking (although it may stimulate it).

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 (2012) has argued previously, sources of big geosocial data are inherently biased toward 'outliers'. In other words, no matter how many geocoded tweets one is able to collect and analyze, they remain limited in their explanatory value for many purposes, as the number of geocoded tweets is but a small fraction[3] 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 toward 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; Kitchin and Dodge 2011).

In order to overcome these limitations, we propose a more systematic methodological and conceptual approach to the geoweb, highlighted in the following section, 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, and 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.

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

Downloaded by [Jeremy Crampton] at 05:18 19 April 2013

the NCAA championship game – began in the late evening of 2 April 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 in the weekend resulted in a sharp peak of chatter on the Lexington Police Department (hereafter, LPD) radio, which is accessible to the public 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 article, 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 data set 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.

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. Geocodable tweets are binned together in 30 square kilometer hexagons in order to more effectively show the varying tweet density in different locations. Data classification is done using a rounded Jenks natural breaks method. 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 outward, especially toward larger cities nearby such as Nashville, Tennessee.

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 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% of all tweets can ultimately be associated with a physical location with some degree of confidence (Graham, Hale, and Gaffney Forthcoming). 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.

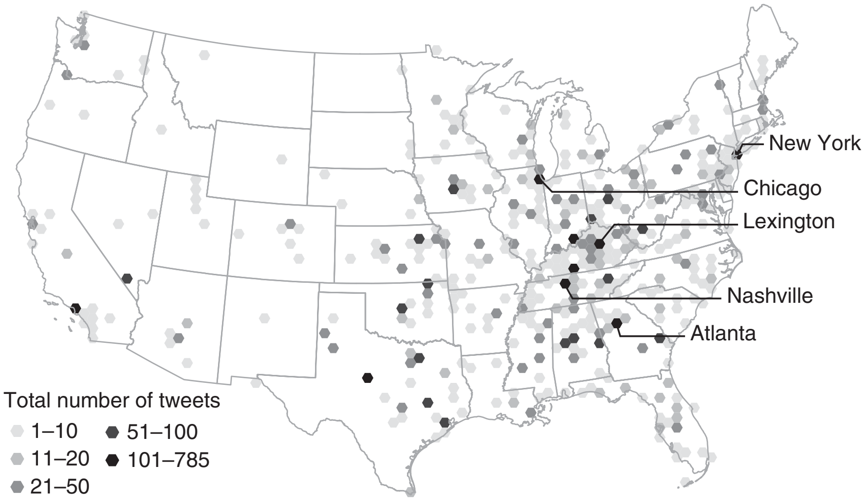


Figure 2. Distribution of all #LexingtonPoliceScanner tweets.

Downloaded by [Jeremy Crampton] at 05:18 19 April 2013

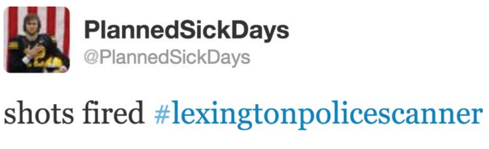
An alternative method for locating geosocial media is the information associated with the actual tweet itself. Derived from GPS coordinates in a mobile device or triangulation from cell or Wi-Fi signals, a geocoded tweet provides the site where the act of tweeting occurs. While this provides a great deal of 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 1% 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% of the 12,590 tweets in the original database were geocodable by user-defined location information in profiles at a minimum of the city level, while only 0.2% 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, data set.

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, 2 April, at 11:50:49 pm, a Twitter user transcribed an audio clip of the police scanner noting that shots had been fired, adding the hashtag #LPS (see Figure 3). 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 the globe, jumping scales from a local occurrence to a worldwide phenomenon and briefly becoming a globally trending topic on Twitter.

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 illustrates how the spatial patterns in tweeting differ from over time. Similar to Figure 2, tweets are binned together in 30 square kilometer hexagons and data are classified based on the aggregate counts for the entire time frame using a rounded Jenks natural breaks method. 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 United States (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).



PlannedSickDays
@PlannedSickDays

shots fired #lexingtonpolicescanner

Figure 3. 'Shots fired' tweet.

In addition to changing spatial extent of #LPS tweeting, Figure 5 highlights how the frequency of these tweets (aggregated in 5-minute bins) 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.

For example, the line reflecting the number of retweets[4] (forwarding a received tweet rather than creating new content) with 60% of the data set 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 insight, as the line representing the frequency of tweets within a 20 mile radius around Lexington (roughly representing Lexington's metropolitan area) exhibits a different pattern than the general trend which corresponds to the findings of Figure 4. Attention to the police scanner peaked about one hour later in Lexington than elsewhere, but the number of tweets generated within Lexington relative to the total number of tweets increased after 3 am – from ~5% at 2 am to ~15% between 3 and 6 am. 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.

Beyond the proximate

An additional dimension to the geographies of #LPS, beyond the changing pattern of tweets over space and

Downloaded by [Jeremy Crampton] at 05:18 19 April 2013

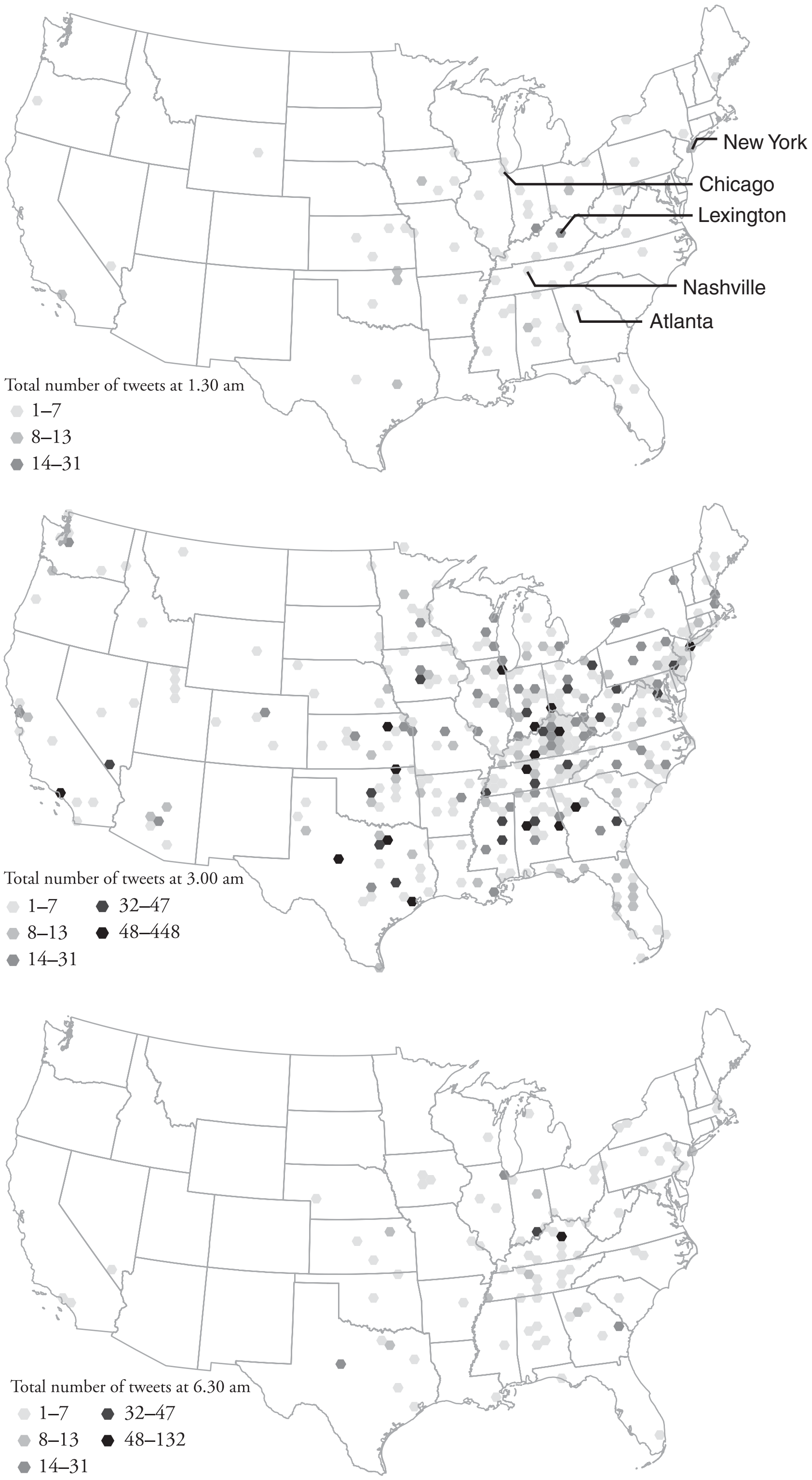

New York
Chicago
Lexington
Nashville
Atlanta
Total number of tweets at 1.30 am
1–7
8–13
14–31
Total number of tweets at 3.00 am
1–7
8–13
14–31
32–47
48–448
Total number of tweets at 6.30 am
1–7
8–13
14–31
32–47
48–132

Figure 4. The geographic distribution of tweets.

time, is the relational connections that collectively constitute the data set; 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

Downloaded by [Jeremy Crampton] at 05:18 19 April 2013

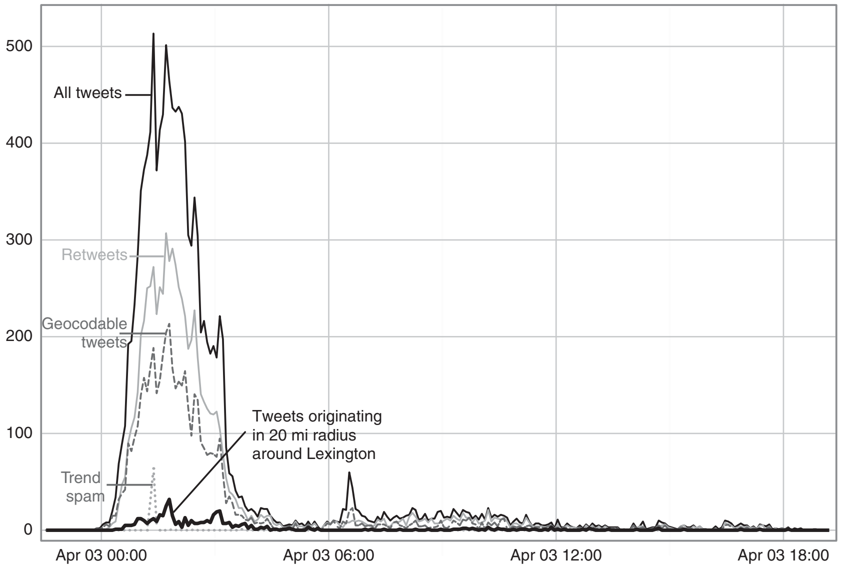


Figure 5. The frequency of tweets over time.

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. A connection, or an edge, means that information (in this case a tweet) flowed between those two users and locations. 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.

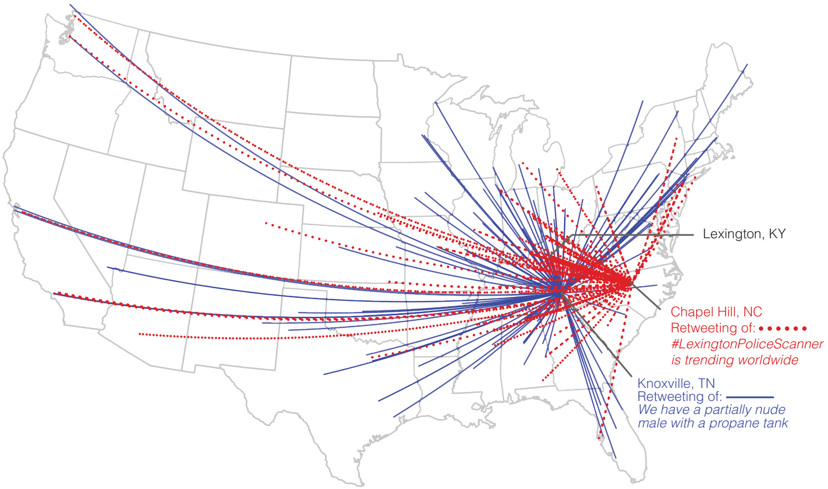


Figure 6. Retweet networks overlaid onto physical space.

Downloaded by [Jeremy Crampton] at 05:18 19 April 2013

While this is only a simple visual example, it 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.

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 data set, 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 single 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 1:18 am: '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 .' Although we handpicked just one example of a bot to show the impact of even a single nonhuman Twitter user, an algorithmic approach could possibly yield a more exhaustive classification of who is human, bot or cyborg (cf. Chu et al. 2010; Grier et al. 2010).

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 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.

Beyond the user-generated

Following the need to understand how nonhuman 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 2012), it is important to look beyond and leverage these sources, even when they 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 data sets, 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 source. Indeed, we would argue that it is absolutely necessary to leverage user-generated data with ancillary data sets if one is to maximize the utility of the data.

As we have seen, following the very first tweet with the #LPS hashtag at 11:50:49 pm, 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 yield only four crime incidents at State Street (the

Downloaded by [Jeremy Crampton] at 05:18 19 April 2013

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 6:30 am 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 provides an 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.[5]

Conclusion

The goal of this article 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 neither 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 have sought to establish a forward-looking program for geoweb research that builds on, rather than merely attacks, existing work in this area. While there remain some significant shortcomings in using such data sources (Haklay 2012, 2013), we believe that the methodological and conceptual program identified here offers some preliminary avenues for overcoming these issues.

And yet we remain cautious of the potential of this research, as broader social, political, economic, and institutional forces remain important in structuring how big data relates to the world that it supposedly describes. At once, such massive data sources, are already under the dual, and often interrelated, threats of commodification by private corporations and surveillance by government intelligence agencies. While we obviously eschew such nefarious motives when using such data sources for academic research, many of the methods employed in this article, such as social network analysis of retweeting patterns shown above, are already being used to disambiguate and identify opinion leaders within various groups, whether to more successfully market particular products or to track potential terrorist threats. As such, we also see significant potential not just in using big data as a source of information on which to construct analysis, but also in studying the ways that big data is embedded in particular social and institutional configurations and employed to achieve particular, and not always benign, ends.

Acknowledgments

This research received funding support for DOLLY through The New Mappings Collaboratory, an initiative of the College of Arts & Sciences at the University of Kentucky.

Notes

[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.

[2] The total number of tweets reached about 170 billion as of January 2013 (Library of Congress 2013).

[3] Based on our database of Twitter activity since December 2011, we estimate that geotagged tweets account for approximately 1.5% of all tweets, though this number is steadily growing.

[4] In this context, retweets are only those tweets that start with 'RT.'

[5] 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 involved fans camping out in tents on the campus in order to get tickets for the season's first practice. From

the imagery, we can estimate the crowd size. We recognize that the issue of unmanned autonomous vehicles (UAVs) or drones is a contentious one, especially given the increasing usage of drones within the United States. Although space precludes it here, an analysis of the political economy of drones and surveillance would illuminate the larger context of this aspect of the geoweb.

References

Amin, A. 2002. "Spatialities of Globalisation." *Environment and Planning A* 34 (3): 385–399.

Anderson, C. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired Magazine* 16: 7.

Boyd, D., and K. Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–679.

Chu, Z., S. Gianvecchio, H. Wang and S. Jajodia. 2010. "Who is Tweeting on Twitter." In *26th Annual Computer Security Applications Conference*, 21. New York: ACM Press.

Crampton, J. 2008. "Cartography: Maps 2.0." *Progress in Human Geography* 33 (1): 91–100.

Elwood, S. 2008. "Volunteered Geographic Information: Future Research Directions Motivated by Critical, Participatory, and Feminist GIS." *GeoJournal* 72 (3–4): 173–183.

Elwood, S. 2010. "Geographic Information Science: Emerging Research on the Societal Implications of the Geospatial Web." *Progress in Human Geography* 34 (3): 349–357.

Elwood, S. 2011. "Geographic Information Science: Visualization, Visual Methods, and the Geoweb." *Progress in Human Geography* 35 (3): 401–408.

Goodchild, M. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69 (4): 211–221.

Gould, P. 1981. "Letting the Data Speak for Themselves." *Annals of the Association of American Geographers* 71 (2): 166–176.

Graham, M. 2011. "Wiki Space: Palimpsests and the Politics of Exclusion." In *Critical Point of View: A Wikipedia Reader*, edited by G. Lovink, and N. Tkacz, 269–282. Amsterdam: Institute of Network Cultures.

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

Graham, M., S. Hale, and D. Gaffney. Forthcoming. "Where in the World Are You? Geolocation and Language Identification in Twitter." *The Professional Geographer*.

Graham, M., S. A. Hale, and M. Stephens. 2011. *Geographies of the World's Knowledge*. London: Convoco! Edition.

Graham, M., T. Shelton, and M. Zook. 2013. "Mapping Zombies". In *Zombies in the Academy*, edited by A. Whelan, C. Moore, and R. Walker. Brighton: Intellect Press

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

Graham, M., and M. Zook. 2013. "Augmented Realities and Uneven Geographies: Exploring the Geo-Linguistic Contours of the Web." *Environment and Planning A* 45 (1): 77–99.

Grier, C., et al. 2010. "@spam." In *17th ACM Conference*, 27. New York: ACM Press.

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, New York, February 25, 2012.

Haklay, M. 2013. "Neogeography and the Delusion of Democratisation." *Environment and Planning A* 45 (1): 55–69.

Hollenstein, L., and R. 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, R., and M. Dodge. 2011. *Code/Space: Software and Everyday Life*. Cambridge, MA: The MIT Press.

Laird, S. 2012. "#LexingtonPoliceScanner: Twitter Listens, Reacts to Kentucky Riots". *Mashable.com*, 3 April 2012. Accessed September 27, 2012. http://mashable.com/2012/04/03/lexingtonpolicescanner-twitter-listens-reacts-to-kentucky-riots-pics/

Laney, D. 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety. [Gartner Group]." Accessed December 18, 2012. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lefèbvre, H. 1991. *The Production of Space*. Malden, MA: Blackwell.

Leszczynski, A. 2012. "Situating the Geoweb in Political Economy." *Progress in Human Geography* 36 (1): 72–89.

Library of Congress. 2013. Update on the Twitter Archive at the Library of Congress. Accessed January 8, 2013. http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf

Massey, D. 1991. "Global Sense of Place." *Marxism Today*, June 1991.

Massey, D. 1993. "Power-Geometry and a Progressive Sense of Place." In *Mapping the Futures: Local Cultures, Global Change*, edited by J. Bird, B. Curtis, T. Putnam, G. Robertson, and L. Tickner, 59–69. London: Routledge.

O'Reilly, T. 2005. "What Is Web 2.0." *O'Reilly Media*. http://oreilly.com/pub/a/web2/archive/what-is-web-20.html

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

Shelton, T., M. Zook, and M. Graham. 2012. "The Technology of Religion: Mapping Religious Cyberscapes." *The Professional Geographer* 64 (4): 602–617.

Wall, M., and T. Kirdnark. 2012. "Online Maps and Minorities: Geotagging Thailand's Muslims." *New Media & Society* 14 (4): 701–716.

Warf, B., and D. Sui. 2010. "From GIS to Neogeography: Ontological Implications and Theories of Truth." *Annals of GIS* 16 (4): 197–209.

Watkins, D. 2012. "Digital Facets of Place: Flickr's Mappings of the U.S.-Mexico Borderlands." Unpublished MA thesis, University of Oregon Department of Geography.

Wilson, M. W. 2011a. "'Training the Eye': Formation of the Geocoding Subject." *Social & Cultural Geography* 12 (4): 357–376.

Wilson, M. W. 2011b. "Data Matter(s): Legitimacy, Coding, and Qualifications-of-Life." *Environment and Planning D: Society and Space* 29 (5): 857–872.

Zook, M., and M. Dodge. 2009. "Mapping, Cyberspace." In *International Encyclopedia of Human Geography*, Vol. VI, edited by R. Kitchin, and N. Thrift, 356–367. Oxford: Elsevier.

Zook, M., and M. Graham. 2010. "Featured Graphic: The Virtual 'Bible Belt'." *Environment and Planning A* 42 (4): 763–764.

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

Downloaded by [Jeremy Crampton] at 05:18 19 April 2013