



HOME ABOUT LOGIN REGISTER SEARCH CURRENT
 ARCHIVES ANNOUNCEMENTS SUBMISSIONS

Home > Volume 18, Number 5 - 6 May 2013 > Leetaru



Mapping the global Twitter heartbeat: The geography of Twitter

by Kalev H. Leetaru, Shaowen Wang,
 Guofeng Cao, Anand Padmanabhan,
 and Eric Shook

Abstract

In just under seven years, Twitter has grown to count nearly three percent of the entire global population among its active users who have sent more than 170 billion 140-character messages. Today the service plays such a significant role in American culture that the Library of Congress has assembled a permanent archive of the site back to its first tweet, updated daily. With its open API, Twitter has become one of the most popular data sources for social research, yet the majority of the literature has focused on it as a text or network graph source, with only limited efforts to date focusing exclusively on the geography of Twitter, assessing the various sources of geographic information on the service and their accuracy. More than three percent of all tweets are found to have native location information available, while a naive geocoder based on a simple major cities gazetteer and relying on the user-provided Location and Profile fields is able to geolocate more than a third of all tweets with high accuracy when measured against the GPS-based baseline. Geographic proximity is found to play a minimal role both in who users communicate with and what they communicate about, providing evidence that social media is shifting the communicative landscape.

Contents

[Introduction](#)
[The native geography of Twitter: Georeferenced tweets](#)
[The linguistic geography of Twitter](#)
[From text to maps: The textual geography of Twitter](#)
[Accuracy and language](#)
[The geography of communication on Twitter](#)
[The geography of linking discourse](#)
[User profile links](#)
[Twitter versus mainstream news media](#)
[Twitter's geography of growth and impact](#)
[Conclusions](#)

[OPEN JOURNAL SYSTEMS](#)

[Journal Help](#)

USER

Username

Password

Remember me

JOURNAL CONTENT

Search

All

Browse

- [By Issue](#)
- [By Author](#)
- [By Title](#)
- [Other Journals](#)

FONT SIZE

CURRENT ISSUE

ATOM	1.0
RSS	2.0
RSS	1.0

ARTICLE TOOLS

 [Abstract](#)

 [Print this](#)

[article](#)

 [Indexing](#)


[metadata](#)

 [How to cite](#)


[item](#)



[Supplementary files](#)

 Email this

article (Login required)

 Email the

author (Login)

Introduction

Since its founding in 2006, Twitter has grown at an exponential rate, today counting among its active users more than 2.9 percent of all people living on Earth (Fiegerman, 2012) and 9.1 percent of the population of the United States and “has become the pulse of a planet-wide news organism, hosting the dialogue about everything from the Arab Spring to celebrity deaths” (Stone, 2012). In its advertising materials, Twitter calls itself “the global town square — the place where people around the globe go to find out what’s happening right now” and that it is “increasingly the pulse of the planet” (Twitter, 2013a). In just the past 12 months alone Twitter has doubled from 100 million to 200 million active users (Twitter, 2011; Fiegerman, 2012), while over the last seven years, more than 170 billion tweets totaling 133 terabytes have been sent, 149 billion of them in just the last 24 months (Library of Congress, 2013).

Twitter offers “an unprecedented opportunity to study human communication and social networks,” (Miller, 2011) while the rising role of Twitter in the consumption of traditional media like television even led Nielsen to create a new Twitter “social TV” rating system (Shih, 2012). During major disasters, governments are increasingly turning to Twitter to provide realtime official information streams and directives (Griggs, 2012), while emergency services are taking the first steps to monitor Twitter as a parallel 911 system, especially in cases where traditional phone service is unavailable (Khorram, 2012). Yet, perhaps the greatest indication of Twitter’s cultural significance is that the Library of Congress now maintains a permanent historical archive going back to the site’s founding and updated daily (Library of Congress, 2013).

Unlike most social network sites, Twitter and its redistributors make nearly all of its data available via APIs that enables realtime programmatic access to its massive seven-year archive. This availability and ease of use has made Twitter one of the most popular data sources for studying social communication, with Google Scholar listing more than 3.2 million papers mentioning the service. Yet, the majority of the literature that has studied Twitter has focused on the text of the tweets or the network graph connecting users (Miller, 2011). Few studies make use of the geographic information attached to tweets, while papers like Poblete, *et al.* (2011) have used it primarily as a filtering mechanism rather than focusing on the geography itself. Most have relied either on natively georeferenced tweets or passing the user’s self-reported location to the Google Geocoder or Yahoo! Placemaker API. Others, like Takhteyev, *et al.* (2012), have integrated geography more closely into their analyses, but have limited themselves to just a few thousand tweets out of the 170 billion sent to date.

More critically, Takhteyev, *et al.* (2012), like most studies that have attempted to geolocate tweets, have relied on the user Location field, making the assumption that it yields the most accurate reflection of a user’s geographic position. Social media monitoring companies, like SemioCast (2012), each use their own proprietary technology to locate tweets geographically, but offer no detail on the data fields or algorithms they rely on or estimates of the accuracy of their approaches. In fact, no major study to date has focused exclusively on the geography of Twitter, examining all available sources of geographic information in the Twitter stream and assessing their accuracy. As location is playing an increasing role in everything from monitoring for natural disasters (Earle, *et al.*, 2011) to “the first ever official United Nations crisis map entirely based on data collected from social media” (Meier, 2012), there is a critical need to better understand the geography of Twitter.

In Fall 2012, supercomputing manufacturer Silicon Graphics International (SGI), the University of Illinois, and social media data vendor GNIP collaborated to create the “Global Twitter Heartbeat” project (<http://www.sgi.com/go/twitter>) in order to map global emotion expressed on Twitter in realtime. GNIP provided access to the Twitter Decahose, which consists of 10 percent of all tweets sent globally each day, while SGI provided access to one of its new UV2000 supercomputers with 256 processors and 4TB of RAM running the Linux operating system. The result of this collaboration, which debuted at the annual Supercomputing 2012 conference held in Salt Lake City, Utah, processed the Twitter Decahose in

required)

ABOUT THE AUTHORS

Kalev Leetaru
<http://www.kalevleetaru.com>
 University of Illinois
 at Urbana-
 Champaign
 United States

Kalev Leetaru is a University Fellow at the University of Illinois Graduate School of Library and Information Science. His work focuses on ?big data? analyses of space, time, and network structure to understand human society and culture through its communicative footprints and is the author of *Data Mining Methods for the Content Analyst* from Routledge.

.....
Shaowen Wang
 University of Illinois
 at Urbana-
 Champaign
 United States

Shaowen Wang is an Associate Professor in the Department of Geography and Geographic Information Science, and founding Director of the CyberInfrastructure and Geospatial Information Laboratory at the University of Illinois at Urbana-Champaign.

.....
Guofeng Cao
 University of Illinois
 at Urbana-
 Champaign
 United States

Guofeng Cao is a Postdoctoral Research Associate in the Department of Geography and Geographic Information Science

real-time, producing a sequence of heatmaps once per second visualizing global emotion and displayed on an 80" LCD monitor and a large 12'-diameter inflatable sphere known as a PufferSphere. In order to construct these visualizations, the project needed to construct one of the most detailed geographic representations of Twitter ever created in order to assign a geographic location to every tweet possible and to understand the geographic biases in the Twitter stream that could affect its results, the results of which form the basis of this paper.

From 12:01AM 23 October 2012 through 11:59PM 30 November 2012, the Twitter Decahose from GNIP streamed 1,535,929,521 tweets from 71,273,997 unique users, averaging 38 million tweets from 13.7 million users each day. The JSON file format in which the stream is encoded generated just over 2.8TB of data over this 39 day period, but the majority of this consists of metadata, with the actual total tweet text weighing in at 112.7GB, containing over 14.3 billion words. The average tweet is 74 characters long and consists of 9.4 words. In all, this dataset encompasses just over 0.9 percent of all tweets ever sent since the debut of Twitter and 35.6 percent of all active users as of December 2012.

Figure 1 shows the total number of tweets received per day from the Decahose over this period, while Figure 2 shows the average number of tweets received per hour. Twitter exhibits strong temporal change, from a low of just over one million tweets per hour from midnight to 2AM PST to just over two million from 7-9AM PST. Twitter's content stream is dominated by a small number of users. The top 15 percent of users account for 85 percent of all tweets, while the top five percent of all users account for 48 percent of all tweets and the top one percent of all users (just 720,365) account for 20 percent of all tweets. A very small number of core users thus drive the majority of Twitter's traffic. A quarter of users active during this period tweeted just once, while half tweeted between one and four times. Roughly 30 percent of users were active a single day (sending one or more tweets that day), while half were active one-three days, and 75 percent of users were active 10 days or less. The top 10 percent of users were active 24-39 days, with about one percent of users active all 39 days.

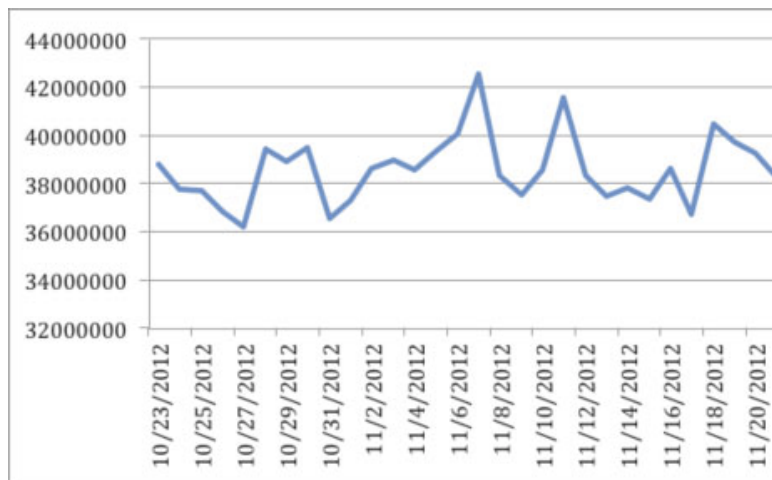


Figure 1: Total tweets per day in the Twitter Decahose 23 October 2012 to 3

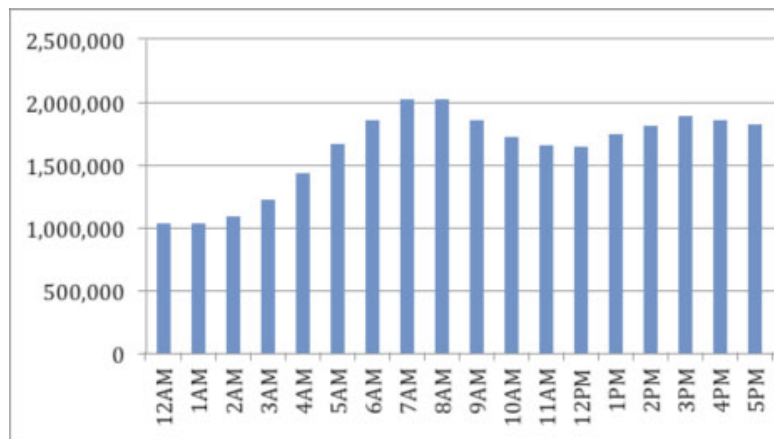
and the CyberInfrastructure and Geospatial Information Laboratory at the University of Illinois at Urbana-Champaign. His research interests include geographic information science, spatiotemporal analysis and location-based social media data analysis.

Anand Padmanabhan University of Illinois at Urbana-Champaign United States

Anand Padmanabhan is a research scientist in the CyberInfrastructure and Geospatial Information Laboratory at the University of Illinois at Urbana-Champaign. His primary research interests include the development of large-scale distributed environments based on Grid computing, multi-agent systems, and P2P systems.

Eric Shook University of Illinois at Urbana-Champaign United States

Eric Shook is a Ph.D. candidate in the Department of Geography and Geographic Information Science and the CyberInfrastructure and Geospatial Information Laboratory at the University of Illinois at Urbana-Champaign. His research interests include large scale agent-based modeling and high performance



computing.

Figure 2: Average tweets per hour in the Twitter Decahose 23 October 2012 to (Pacific Standard Time zone).

The native geography of Twitter: Georeferenced tweets

Since August 2009, Twitter has allowed tweets to include geographic metadata indicating the location where the tweet was authored (Twitter, 2009). There are two types of tweet geolocation information available: Place, which allows a user to manually specify a city or neighborhood using a software menu, and Exact Location, which is a set of coordinates usually provided via GPS or cellular triangulation (Twitter, 2013b). Place locations must be manually selected by the user from a predefined list of locations supported by Twitter, and is primarily used when tweeting from a desktop or fixed-location device. This location must be manually updated by a user, so tweets from the user traveling to another country would still reflect his or her last selected location (Twitter, 2013c). In contrast, Exact Location uses a mobile device's geolocation features to provide the user's geographic location at the time each tweet is sent, meaning the user does not have to take any action to update her location as she travels. Exact Location tweets reveal the user's current location to four decimals, meaning it can capture a precise street address such as a house or favorite coffee shop, associating significant privacy risks with its use (Twitter, 2013c). Due to these privacy risks, tweet geolocation is disabled by default and users must explicitly alter their account settings to enable it. On a typical day during the period studied, 2.02 percent of all tweets included geographic metadata, with 1.8 percent having a Place indicator, 1.6 percent having Exact Location, and 1.4 percent having both (these sum to more than the total because tweets can have both).

Both Place and Exact Location geographic metadata appear in a dedicated "Geo" metadata field designed for automated processing. However, a close inspection of the Decahose shows that an additional 1.1 percent of all tweets instead provide Exact Location coordinates in the user-defined Location field. This textual metadata field is traditionally used to manually enter location in textual form for display on a user's Twitter profile page and is not processed or validated by Twitter. Nearly all tweets with Exact Location coordinates in their Location field have blank Geo metadata fields, meaning these tweets are invisible to Twitter mapping systems that look only at the Geo field. The presence of Exact Location coordinates in the Location field has not been extensively explored in the literature and appears to result primarily from iPhone and BlackBerry-based Twitter clients. When these tweets are combined with those from the Geo metadata field, 46,672,798 tweets, or 3.04 percent of the Decahose, was georeferenced, capturing an average of 600,000 unique points on Earth each day.

In all, a total of 16,098,212 distinct locations were recorded in those

georeferenced tweets, with 15,909,111 unique Exact Location values and 196,843 distinct Place values. (These sum to slightly more than the total because some coordinates were present in both Exact Location and Place fields.) The low number of distinct Place values, despite being present in 1.8 percent of all tweets, is due to the fact that the Place field captures location only at the city level, meaning that all users in New York City would have the same value, whereas with Exact Location, even a stationary user would likely report a different location with each tweet due to imprecision and jitter in GPS and cellular triangulation data. [Figure 3](#) displays the geography of the Place field, illustrating that it is limited to a small number of countries. The United States, Canada, Mexico and Puerto Rico all have strong coverage, but the only country in South America with Place coverage is Brazil, while in Africa just South Africa and Morocco have entries. Western Europe is covered, but there is little coverage in Eastern Europe and minimal coverage in Russia. In Asia just Malaysia and Indonesia have coverage. This suggests that while the Place field increases the total number of georeferenced tweets by more than half a percent, it may skew the data towards certain countries. [Figure 4](#) shows all Exact Location coordinates, illustrating that when incorporating sensor-based location information, Twitter exhibits strong geographic diversity, with most countries having at least some georeferenced tweets.



Figure 3: All Place coordinates in the Twitter Decahose 23 October 2012 to 30 November 2012. For a higher resolution version of this figure, go to <http://www.sgi.com/go/twitter/ima> and for a very high-resolution version at <http://www.sgi.com/go/twitter/images/hires/f>



Figure 4: All Exact Location coordinates in the Twitter Decahose 23 October 2012
To see a higher resolution version of this figure, go to <http://www.sqi.com/go/figure4.png>; very high-resolution version at <http://www.sqi.com/go/twitter/ir-highres.png>.

This map exhibits remarkable detail, tracing major road and transportation networks and demonstrating the ability of Twitter on mobile devices to trace society's daily life. A close inspection of this image will immediately prompt comparison with the NASA Visible Earth City Lights imagery (NASA, 2000), which maps the presence of electric lighting at night across Earth, as measured by satellite. The NASA imagery measures urbanization and electrification, indicating areas more likely to have Internet access. [Figure 5](#) overlays the locations of all georeferenced tweets (combining [Figure 3](#) and [Figure 4](#)) on top of the NASA Earth City Lights satellite image, coloring tweets red and night lights blue. This results in a composite image in which bright white areas are those with an equal balance of tweets and electricity, while red areas have a higher density of tweets than night lights and blue areas have more night lights than tweets. Iran and China show substantially fewer tweets than their electricity levels would suggest, reflecting their bans on Twitter, while India shows strong clustering of Twitter usage along the coast and its northern border, even as electricity use is far more balanced throughout the country. Russia shows more electricity usage in its eastern half than Twitter usage, while most countries show far more Twitter usage than electricity would suggest.

To quantitatively measure their similarity, both images were divided into a 1x1 degree latitude/longitude grid (rounding off the fractional portion of their coordinates), resulting in a 180x360 grid with 64,800 cells. The number of tweets and the number of lit pixels in the NASA image were tallied for each grid cell and the similarity of the two resulting grids was tested using a Pearson correlation. The volume of tweets and the penetration of electricity were found to be correlated at $r=0.79$, indicating very high similarity. Intuitively, this makes sense in that Twitter is more likely to be used as a part of daily life in areas that have readily available electricity to support the landline or mobile Internet connectivity needed by Twitter. It also demonstrates that despite high mobile use, Twitter is not a replacement for satellite and other air and space-based sensor systems of society — it is still reliant on the same electrical and network infrastructure as other Internet media and thus has difficulty penetrating into rural areas with low availability of electricity. At the same time, the substantial correlation of georeferenced tweets with the ready availability of electricity suggests these tweets are likely to be highly representative of where Twitter users are most likely to be found. Despite less than three percent of all tweets having geolocation information, this suggests they could be used as a dynamic reference baseline to evaluate the accuracy of other methods of geographic recovery.



Figure 5: Comparison of georeferenced tweets from the Twitter Decahose 23 November 2012 with NASA Visible Earth imagery (red areas overemphasize, dark areas underemphasize). To see a higher resolution version of this figure, go to <http://www.firstmonday.org/ojs/index.php/fm/article/view/42111/images/hires/figure5.png>.

While georeferenced tweets may be evenly distributed geographically according to the availability of electricity, they are generated by only a small portion of Twitter's userbase. Just 8.2 percent of all users active during this period had either Place or Exact Location information available for their tweets, with 4.2 percent of all users sending a single georeferenced tweet, accounting for 9.7 percent of all georeferenced tweets. Just one percent of all users (722,692 in the Decahose) accounted for 66 percent of all georeferenced tweets, indicating georeferenced tweets are created by an even more extreme subset of users than overall tweets. Given that two-thirds of the native geography of Twitter is driven by just one percent of all users, this suggests that studies relying purely on these tweets will have a skewed view of the Twitterverse, especially over short periods of time.

There is no measurable difference in the density of georeferenced tweets between weekdays and weekends. This suggests that the geography of Twitter is driven primarily by the subset of users who have geolocation turned on for their account, rather than a difference in communicative behavior between the workweek and weekends. Given that geolocation must be enabled or disabled for a user's entire account, rather than toggled on a per-tweet basis, it makes sense that users would leave it enabled or disabled at all times. It also suggests that users use the same client to tweet at all times, rather than using a fixed desktop during the workday and a GPS-enabled mobile device during the weekend. [Figure 6](#) shows that while there is not a substantial difference in georeferencing by day of the week, there is a difference by time of day, with georeferenced tweets using the Geo metadata field peaking at 2.3 percent of all tweets at 1PM PST through a low of 1.7 percent at 6AM PST. Reflecting back to the earlier maps, this is likely more reflective of the differing penetration of geolocation across the world than suggestive of a difference in how users report their location over the course of a day.

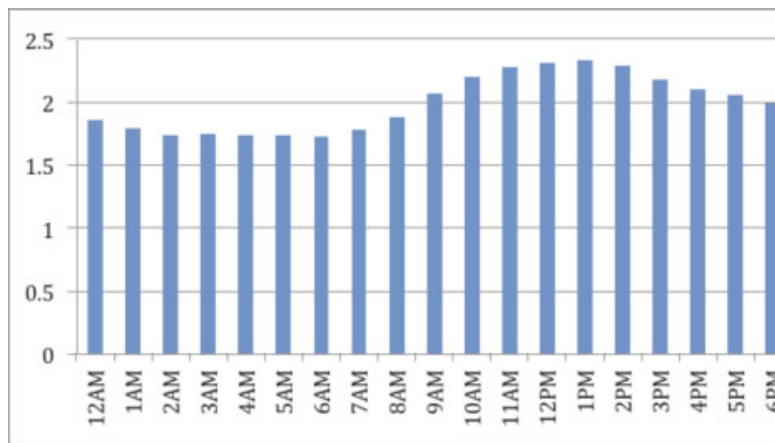


Figure 6: Percent of Twitter Decahose tweets 23 October 2012 to 30 November 2012 georeferenced (PST).

[Table 1](#) examines this more closely by ranking all cities globally by the total percentage of all georeferenced tweets originating from that city. Since Exact Location tweets are specified in latitude/longitude coordinates, rather than the name of a city, the centroid of all cities on earth with a population of more than one million was compiled and all tweets within one degree latitude/longitude of each centroid were tallied. The strong presence of Twitter in the United States is reflected in the fact that six of the top 20 cities are from the United States. Jakarta alone accounts for nearly three percent of all georeferenced tweets, illustrating Indonesia’s outsized presence on Twitter, while New York City and São Paulo are nearly tied for second. Texas stands out in that two cities, Dallas and Houston, both make the top 20 list, with a third city, San Antonio, at number 42, with 0.32 percent.

City	Percentage georeferenced tweets
Jakarta	2.86
New York City	2.65
São Paulo	2.62
Kuala Lumpur	2.10
Paris	2.03
Istanbul	1.60
London	1.57
Rio de Janeiro	1.39
Chicago	1.28
Madrid	1.17
Los Angeles	1.14
Singapore	1.05
Houston	1.04
Mexico City	1.03
Philadelphia	0.99
Dallas	0.91

Manila	0.90
Brussels	0.88
Tokyo	0.85
Moscow	0.77



The linguistic geography of Twitter

Georeferenced tweets have the distinct advantage that their location information is provided in native geographic format, making them language agnostic. However, the extremely low proportion of such tweets, comprising just over three percent of all tweets, means that to expand the universe of mappable tweets, geocoding algorithms must be used to identify and disambiguate textual mentions of place, such as converting a mention of "New York City" in a tweet into a set of mappable coordinates (Leetaru, 2012). Most geocoding algorithms, however, have been designed only for English text. The linguistic geography of Twitter is therefore critical: if English is rarely used outside of the United States, or if English tweets have a fundamentally different geographic profile than other languages outside of the United States, this will significantly skew geocoding results.

In 2011, Eric Fischer applied the Google Chrome language detection library to automatically determine the language of each georeferenced tweet in the Twitter streaming API from 14 May to 20 October 2011 to map the geographic distribution of language on Twitter. GNIP uses its own algorithms to determine the language of each tweet in the Twitter Decahose and recognizes 26 major languages, shown in [Table 2](#). The first column shows the percent of all georeferenced tweets that were published in that language, indicating its prevalence in the native geography of Twitter, the second column shows what percent of all tweets, georeferenced or not, were published in that language, and the third column shows what percent of all tweets in that language were georeferenced. English is by far the most common language on Twitter, accounting for 38.25 percent of all tweets and 41.57 percent of georeferenced tweets. Yet, just 2.17 percent of all English tweets are georeferenced, indicating that the vast majority of tweets in the language do not carry native geographic information. Spanish is the second most popular georeferenced language at just a quarter of English, but for georeferenced tweets, it is tied with Japanese. Just 11 languages have more than one percent each of georeferenced tweets. Three languages exhibit substantial differences in their use between georeferenced and non-georeferenced tweets: Japanese moves from eleventh to second most popular, Norwegian moves from thirteenth to fifth, and Korean moves from nineteenth to twelfth. GNIP's language detection engine assigns a value for all but 0.51 percent of all tweets, but eight percent of georeferenced tweets do not have a language assignment. The fact that nearly one third of tweets without a language assignment are georeferenced suggests that georeferenced tweets may have a higher density of hyperlinks, especially tweets that contain only a link with no additional text and thus cannot be classified by language.

Table 2: Percent georeferenced tweets by language (Twitter Decahose 23 October 2012 to 30 November 2012).

	Percentage georeferenced tweets	Percentage all tweets	Percentage language georeferenced
English	41.57	38.25	2.17
Spanish	11.16	11.37	1.96
Portuguese	9.50	5.58	3.40
Other	8.39	0.51	32.78

Indonesian	7.33	8.84	1.66
Turkish	3.87	1.80	4.29
French	3.85	2.30	3.35
Arabic	2.81	4.09	1.37
Russian	2.24	1.12	3.98
Italian	1.95	1.31	2.97
Japanese	1.63	11.84	0.27
Dutch	1.40	1.51	1.85
Norwegian	0.76	7.74	0.20
German	0.75	0.66	2.25
Swedish	0.48	0.27	3.63
Thai	0.46	0.48	1.92
Finnish	0.44	0.34	2.62
Polish	0.40	0.34	2.34
Korean	0.36	1.17	0.62
Czech	0.13	0.11	2.35
Danish	0.13	0.09	2.90
Greek	0.11	0.07	3.20
Chinese	0.11	0.09	2.53
Ukrainian	0.09	0.04	4.14
Vietnamese	0.03	0.04	1.80
Persian	0.02	0.03	1.28
Hebrew	0.01	0.01	2.49

Simply knowing the overall breakdown of tweets by language does not address the spatial distribution that is so critical for mapping: if all of the English tweets are in the United States, English-based geocoding will not be able to cover the rest of the world. In his 2011 work, Fischer developed a color scale for each language that maximized its contrast with neighboring languages. [Figure 7](#) uses Fischer's color scheme to map the spatial profile of each of the 26 languages recognized by GNIP. In cases where multiple languages are present at the same coordinate, the point is assigned to the most prevalent language at that point and colored accordingly. Most countries show strong homogeneity with a single language in predominate use and small isolated pockets of other languages. However, the region between Germany and Greece is extremely multilingual with Hungary and Serbia in particular having no single language that appears to dominate, while Lebanon, Israel, and the West Bank also have a very strong mix of languages. The continent of Africa shows very sparse Twitter usage with a sizable English population centered primarily in South Africa, Kenya, Nigeria, and Ghana, with France's influence visible, especially in Morocco, Algeria, and Tunisia. Thailand is dominated by Thai, but also shows a rich diversity of other languages in use as well. The United States is overwhelmingly English, but shows a strong scattering of other languages throughout the country, especially in the Midwest, while India also appears to make heavy use of English.

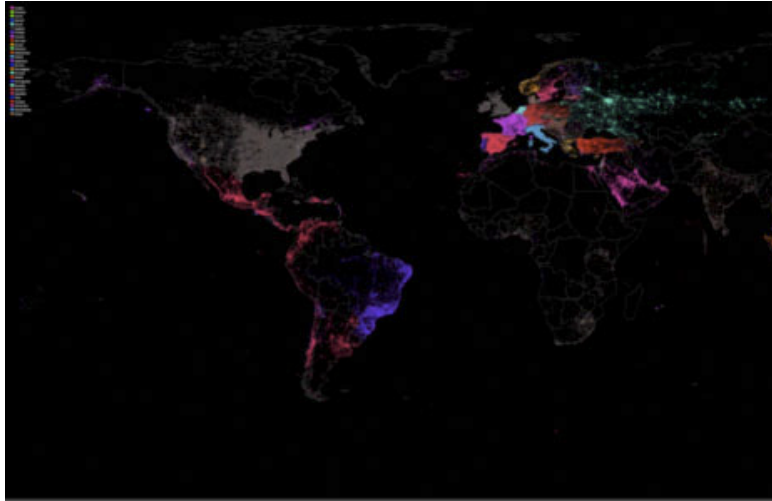


Figure 7: Twitter Decahose georeferenced tweets 23 October 2012 to 30 November 2012 (color scale from Fischer, 2011). To see a higher resolution version <http://www.sqi.com/go/twitter/images/hires/figure7.png>; very high-resolution <http://www.sqi.com/go/twitter/images/hires/figure7-highres.jpg>

Since each location is colored by the language most prevalent there, less-used languages will be drowned out by domestic ones. This makes it difficult to explore the spatial profile of English and how significantly it is used outside of the few English-speaking nations, especially its penetration into countries with strong domestic languages. For example, while the majority of tweets sent in France are in French, do English-language tweets have a similar spatial profile, or are they primarily clustered around major tourist areas? [Figure 8](#) filters the map above to display only English-language tweets, showing all locations where English tweets were sent from, even if there were more tweets from other languages at those locations. While [Figure 7](#) shows that most countries are dominated by their own languages, [Figure 8](#) shows that English is still spoken widely in nearly every country and with a nearly identical geographic distribution.

To quantify the similarity of the geographic profile of English with that of all languages, the world was once again divided into a 1x1 degree grid and the Pearson correlation calculated, measuring for each cell the number of English-language tweets and the number of tweets in any language. This yielded a correlation of $r=0.75$, indicating the two are highly similar. For any location on earth, the relative percent of all English tweets posted from that location is correlated with the relative percent of all tweets of any language posted from that location. This suggests that English offers a spatial proxy for all languages and that a geocoding algorithm which processes only English will still have strong penetration into areas dominated by other languages (though the English tweets may discuss different topics or perspectives).



Figure 8: Twitter Decahose English-language georeferenced tweets 23 October 2012. To see a higher resolution version of this figure, go to <http://www.sgi.com/figure8.png>.

From text to maps: The textual geography of Twitter

The limited availability of natively georeferenced tweets and strong prevalence of English suggests geocoding algorithms could be used to considerably expand the universe of mappable tweets by extracting geographic information from the textual information of the Twitter stream. This raises questions regarding which fields contain the greatest amount of recoverable textual geographic information, which provide the highest accuracy, how to assess that accuracy, and which geocoding algorithms achieve the best results with the limited text available in Twitter.

Of the more than 50 metadata fields provided with each tweet, several are of particular interest to geographic recovery. Perhaps most obvious is the user Location field, which allows users to type their location in text form and is available in 71.4 percent of all tweets. This field is separate from the Exact Location and Place fields and is text-based, meaning users may type anything into it, with no validation check performed by Twitter. Users may also include a biographical Profile for their Twitter account, which 87.2 percent of all tweets have, which may also include geographic information. Approximately 78.4 percent of tweets include the user's time zone in textual format, which offers an approximation of longitude, and 74.9 percent of tweets include a processed version of the time zone that gives the user's numeric offset in seconds from UTC. Just over half of tweets, 54.7 percent, include the name of a city in the time zone field, such as "Chicago," rather than just the name of a time zone such as "Central Standard Time," suggesting the time zone field may have specific geographic information that enables the recovery of both latitude and longitude information. Finally, the text of the tweet itself may mention one or more locations.

In each of these fields other than the numeric time zone, any geographic information in the field is expressed in textual form, such as "New York City," requiring geocoding algorithms to identify the location out of surrounding text, disambiguate it, and convert it to approximate mappable coordinates. There are two major types of geocoding algorithms: traditional geocoding, which operates on text where the entire text is known to be a location, and fulltext geocoding, where it must parse a larger body of text to identify the locations mentioned within (Leetaru, 2012). The Location field is likely to contain only the name of a city or other geographic location and so represents a traditional geocoding task,

while a user profile might mention a geographic location amongst paragraphs of non-geographic text (such as "I am a Computer Science student living in New York City and I really love it here").

Nearly one third of all locations on earth share their name with another location somewhere else on the planet, meaning that a reference to "Urbana" must be disambiguated by a geocoding system to determine which of the 12 cities in the world it might refer to, including 11 cities in the United States with that name (Leetaru, 2012). This disambiguation process is heavily dependent on context and estimations based on global geographic reference patterns and therefore prone to error. One of the greatest challenges with assessing the accuracy of geocoding systems is the lack of gold reference datasets against which their results may be compared. For example, when applying full-text geocoding to a collection of documents, one must ascertain whether the system missed any geographic references, whether it identified potential locations that were not actually locations in the text, and whether it properly disambiguated the locations it did identify. This is usually done with a manual review of a random subset of documents, but humans are extremely slow and error-prone at this process themselves, necessitating very small test samples.

Twitter presents a truly unique opportunity to measure the accuracy of the geocoding process in a far more comprehensive and quantitative fashion. Unlike any other major document collection available for academic research, Twitter contains a sensor-based gold standard in which 2.6 percent of the data is assigned a precise physically-measured location that can be compared with the geocoded results to determine their correlation. Even more significant is the size of the data samples: more than 30 million tweets georeferenced via sensor as a gold standard and a total of 1.5 billion tweets containing 14.3 billion words as the collection to geocode. This appears to be the largest gold standard assessment of a geocoding system ever presented in the literature.

Each source field and geocoding algorithm must be assessed along two dimensions: coverage (what percent of all tweets it was able to recover geographic information from) and accuracy (how closely the final results match the gold standard baseline). One complicating factor in assessing accuracy is that natively georeferenced tweets have precision down to a specific street address, while the output of textual geocoding systems is a city centroid. To enable direct comparison, a 1x1 degree grid was used to compute the Pearson correlation of the output of each geocoding approach against the georeferenced gold standard to assess its accuracy, while the total percent of all tweets recognized by the algorithm was computed to assess its coverage. An algorithm or source field which achieves very high coverage, identifying location mentions across a large number of tweets, but which is a poor match for the gold standard is not extremely useful as a geocoding approach. Similarly, one which matches very few tweets, but has very high accuracy will not fulfill the requirement of expanding the percentage of tweets that may be mapped. The goal, therefore, is to find a combination of source field(s) and algorithm(s) which yield the highest coverage and accuracy simultaneously.

Fulltext geocoder

To provide an initial baseline of accuracy and coverage using a known geocoding algorithm that has been applied to many different document corpuses, the fulltext geocoding system from Leetaru (2012) was explored first. The tweet text, Location, and Profile fields were concatenated with a comma between each field and passed to the fulltext geocoder. This matched 28.38 percent of all tweets and correlated with the georeferenced baseline at $r=0.51$, indicating strong alignment. Each field was then tested individually, with the tweet text by itself yielding a match rate of 2.44 percent of all tweets at $r=0.25$, the Profile matching 4.72 percent of tweets at $r=0.26$, and the Location field matching 23.96 percent of tweets at $r=0.52$. Combining the Location with the tweet text, 25.13 percent of tweets are matched at $r=0.52$ and combining the Location with the Profile yields 26.42 percent at $r=0.52$. Thus, it is clear that for the fulltext geocoder, the majority of matches stem from the Location field and matches incorporating that field are the most accurate. Both the tweet text and Profile fields contain geographic information, but not in substantial quantities and have poor accuracy.

Of particular note is that the difference between using all three fields (tweet text, Location, and Profile) and using just the Location and Profile fields is a reduction of coverage by just 1.95 percent, but an increase in correlation of 0.01, providing further evidence that the additional geographic information captured from the tweet text degrades the accuracy of the geocoding. This could indicate that mentions of locations in the text of tweets refer to locations being discussed by the user rather than locations near her. Most critically, however, while the contents of each tweet is different, the Location and Profile fields likely do not change regularly for most users and thus are largely constant across all tweets sent by that user. This suggests that instead of geocoding tweets, a production system could geocode users instead and construct a lookup table of all usernames to map tweets by that user to his or her current location. In the case of the Twitter Decahose, in a typical month this would mean geocoding just 71M distinct users, rather than 1.5 billion tweets, a reduction in computation of more than 21 times.

While a coverage rate of 28.4 percent indicates that more than a quarter of all tweets are being geocoded, it is still relatively low, especially given that more than 38 percent of all tweets are in English. To test whether the low match rate is simply a result of a lack of geographic information in Twitter, or whether there is an issue with the geocoding algorithm, the gazetteers used by the geocoder were extracted and a simple text search was performed on each tweet to identify the appearance of any entry from the gazetteers, without attempting to determine whether the word was actually being used in the context of a location or attempting to disambiguate it. First, any mention of a country name across all three fields was tallied, resulting in a match rate of 14.72 percent and a correlation of $r=0.09$, while 4.54 percent of tweets mentioned a U.S. state, with a correlation of $r=0.06$. While passing the threshold of statistical significance, these correlations are far too low to use for actual geocoding due to the coarse resolution of country and state centroids compared with the high resolution of the baseline. The very low country-level match rate, just 14.72 percent, suggests that the majority of location references on Twitter are to cities and other local landmarks, rather than country names. This is beneficial, in that it allows for a high level of localization, but also complicates the matching process, in that city names require far more disambiguation than country names.

Next, any match of an entry from the GNS and GNIS gazetteers was tallied (Leetaru, 2012). For the Location field this resulted in a 50.56 percent match rate and for all three fields this resulted in an 88.72 percent match rate. To test correlation, two methods were used. One used only the first gazetteer match from each tweet, in the order of Location, Profile, and tweet text, and the other recorded every match found in the text. For the Location field only, this resulted in $r=0.17$ for all matches and $r=0.13$ for first match only. For all three fields, this resulted in $r=0.08$ for all matches and $r=0.11$ for first match only. Such extremely high match rates indicate that there is indeed substantial geographic information in tweets, though the low correlations indicate that many of these matches are likely false positives or incorrect matches due to the lack of disambiguation. This suggests that while the fulltext geocoder's coverage rate is relatively low, increasing that coverage rate by disabling the disambiguation process reduces accuracy to just above minimal statistical significance.

One of the challenges faced by the fulltext geocoder is that it relies heavily on surrounding context to separate location from non-location references and to disambiguate candidate locations to their appropriate match. Since it is designed to work with large corpuses of text where there are many common words that can also be part of location names, the algorithm is designed to err on the side of exclusion and only match potential locations where it has additional supporting indicators in the text. This allows it to recognize extremely small townships and geographic landmarks from around the world, but places it at a distinct disadvantage when processing the minimal text of Twitter.

Wikipedia gazetteer

Given that there appear to be a substantial number of geographic mentions on Twitter and that most appear to be to specific cities and landmarks, rather than simple country mentions, the next approach was to construct a more extensive gazetteer of major global locations that do not require disambiguation. While the fulltext geocoder already has a database

of just over 1,600 capital and major cities around the world, it only contains entries for cities that are largely unambiguous. For example, most mentions of "Paris" refer to the capital of France unless they contain additional qualifiers such as "Paris, Illinois." Previous work on the geography of Wikipedia (Leetaru, 2012) resulted in a database of 583,414 English-language entries extracted from the encyclopedia that contained one or more hand-assigned geographic coordinates. This was used to create an English gazetteer that contains the title of each Wikipedia entry and the coordinates of the first geotag location found on that page.

One of the greatest advantages of Wikipedia over the GNIS and GNS gazetteers used by the fulltext geocoder is that its vastly fewer entries (just 5.4 percent of the size of GNS and GNIS) suggests it will only contain larger and more "important" cities. For example, GNS lists three different cities in Indonesia with the official name "Tulungagung": one each in Sulawesi Selatan, Jawa Timur, and Sulawesi Barat administrative divisions. In the absence of further information, the fulltext geocoding engine is unable to distinguish among these and thus cannot geolocate that reference. Wikipedia, on the other hand, has only a single entry for a city named "Tulungagung," linking to the city in Jawa Timur. It is unclear whether this is the correct entry for the location mentioned in a given tweet on Twitter, or if Wikipedia has similar coverage for all countries, but it does suggest Wikipedia's more narrow coverage might significantly enhance the coverage and accuracy of the geocoder.

All 583,414 geotagged Wikipedia pages were compiled into a gazetteer, using the first coordinate found on each page to geolocate it. Entries that contained the full name of an administrative division, such as "West Sulawesi Province" or "Champaign County" were modified to drop endings such as "department," "county," and "province." Similar to the GNS/GNIS test, a simple keyword search was used to identify any mentions of an entry from the gazetteer in each tweet. When applied to just the Location field, 55.24 percent of all tweets were matched with a correlation of $r=0.38$. However, when applied to all three fields, the Wikipedia gazetteer method yielded an astonishing 95.33 percent match rate, but a correlation of just $r=0.30$. One reason for the massively high match rate is likely due to the prevalence of common names and words in many location names. For example, in Arkansas, "Cross County" becomes "Cross" after "county" is dropped off the end of the name, while "Columbia County" becomes "Columbia," and "Howard County" and "Drew County" become Howard and Drew, respectively.

The fact that the Wikipedia gazetteer was able to match nearly the entire Twitter corpus, with an accuracy rate that, while not extremely high, was still higher than several of the previous approaches, suggests this method could have merit. In particular, further work cleaning the Wikipedia gazetteer, especially removing common names and words, could dramatically increase its accuracy. In addition, the fact that Wikipedia is available in 285 different languages suggests it could be used to quickly develop a multilingual gazetteer to recognize locations in other languages.

Global major city gazetteer

The potential of the Wikipedia gazetteer led to testing a third approach, of creating a more limited gazetteer of major cities, essentially expanding the Well Known Places list of the fulltext geocoder with a much larger collection of cities. The MaxMind Free World Cities Database (MaxMind, 2013) contains a list of 47,004 cities from around the world, together with their populations and approximate city centroids. This list vastly expands on the fulltext geocoder's Well Known Places database, containing many cities for which there are multiple cities on earth with that name, while still being far smaller than the Wikipedia gazetteer. All city names less than four characters in length or which were in the fulltext geocoder's blacklist database were discarded and the remaining cities sorted by population. In cases where multiple cities had the same name, the one with the largest population was selected. This resulted in a database of 37,929 entries, which was expanded by adding in the fulltext geocoder's Well Known Places database to add localized transliterations of major city names. For example, while the population database includes the popular transliteration Jiddah, the Well Known Places database adds the alternative spellings Jaddah and Jeddah, which are used extensively on Twitter, and similarly adds Makkah for Mecca.

As with the Wikipedia gazetteer, a simple keyword search was used to search for matches in the text. When it was applied to the Location field, this algorithm yielded a 29.81 percent match rate and exhibited an unprecedented correlation of $r=0.72$. Applied to all fields, the coverage rate increased to 36.96 percent, while the correlation remained at $r=0.72$. Examining each field individually, the Profile yields 7.95 percent match at $r=0.59$, the tweet text yields 5.59 percent at $r=0.45$, the Location and tweet text combined yields 33.06 percent at $r=0.72$ and the Location and Profile combined yields 34.05 percent at $r=0.72$. Thus, as before, the Location field appears to offer the most accurate information on the user's actual location, while the Profile and tweet text are less accurate, but not as poor as with previous methods. Combining these fields with the Location yields significantly higher coverage, but no measurable decrease in correlation, likely due to the fact that they have relatively high correlations on their own. As with fulltext geocoding, the difference between using just the Location and Profile together and combining them with the tweet text is a reduction of just 2.92 percent in the match rate, but no decrease in accuracy.

Finally, as noted earlier, the Timezone field contains a high density of city mentions, suggesting it might be a source of additional geographic information. Applying the major city gazetteer algorithm to this field yields a 77.82 percent coverage rate, but a correlation of just $r=0.34$. The low correlation is likely due to the fact that timezone information is used to set a user's longitude and thus many software programs offer just a few of the largest cities in each timezone for the user to pick from, rather than an exhaustive list of all cities in that timezone. In the United States, for example, some programs offer only Chicago for users in Central Standard Time, meaning a user in southern Mississippi would be forced to pick a city nearly a thousand miles away.



Accuracy and language

One of the primary reasons that the naïve major cities geocoder outperforms the fulltext geocoder is the fulltext geocoder's reliance on disambiguation. All of the cities listed in the MaxMind database are in the GNS and GNIS gazetteers used by the fulltext geocoder, but because the fulltext geocoder must also match much smaller and more obscure locations, it will not confirm a match to a location unless it can find additional corroborating information in the text, such as another major city from that same region or country (Leetaru, 2012). For example, a reference to "Cornwall" could easily refer to a person's name, to one of four cities in four different countries (Australia, Canada, Jamaica, or the United Kingdom), or to one of 10 locations in the United States in nine different states. Without additional context, it is simply impossible to know for certain which is the proper parsing of the word. In addition, the fulltext geocoder was designed to recognize mentions of even very small cities, including the following cities which appeared in the Decahose during this period: Chugiak, Alaska; Rotorua, New Zealand; Chittenango, New York; Hummelstown, Pennsylvania; Kebon Sirih, Menteng in Central Jakarta, Indonesia.

At first glance it would make sense that Twitter, being a local information source populated by average citizens, might make heavy use of local location references. However, even if that is the case, the lack of surrounding context makes it impossible for the geocoder to disambiguate those that do appear. The limited space of Twitter means that communications utilize shorthand and rely on shared background knowledge more heavily. The major cities gazetteer approach works around this limitation by removing the disambiguation process through limiting itself to a far smaller subset of locations and skewing its results towards larger cities. This means that all references to "Urbana" will always be coded as mentions of Urbana, Illinois, even if they actually were about Urbana, Ohio, but as the $r=0.72$ correlation demonstrates, this approach appears to be accurate for Twitter. In addition, the elimination of the disambiguation process means the major city gazetteer algorithm is extremely fast, processing 4,510 tweets per second (390M/day) on a single 2.6Ghz processor core, sufficient to handle the full Twitter firehose. Since the most accurate results stem from coding just the Location and

Profile fields, a production system could code just user accounts, rather than tweets, at a loss of coverage of less than three percent of all tweets, and with a computational reduction from geocoding 37M daily tweets in the Dechahose to just 1.3M user accounts per day.

While a correlation of $r=0.72$ is extremely strong, it does indicate that there is not a perfect alignment of the geocoding results with the georeferenced baseline. One potential source of error is the "suburb effect," where users in the suburbs of major metropolises using georeferenced tweets are correctly placed in their neighborhoods outside the city, but geocoded users provide the location of the major city they are closest to. Thus, a user living in a suburb of Chicago might list "Chicago" as his location, which would place him at the centroid of downtown Chicago, while a georeferenced tweet would appear where his house is located, 20 minutes away. To test this, the correlation for the major cities gazetteer applied to the Location and Profile fields was recomputed using a 5x5 degree grid which groups major cities with their suburbs. This increased the correlation up to $r=0.83$, while a 10x10 degree grid, testing whether mentions were in the same general region, increased the correlation to $r=0.89$. Thus, a significant portion of the error lies in the slight misalignment between the high-resolution street-level georeferenced baseline and the coarser city centroid resolution of the geocoding results. This argues that users are largely truthful in providing their locations on Twitter and is in keeping with recent research suggesting that social media like Twitter (Castillo, *et al.*, 2013) and Facebook (Back, *et al.*, 2010) may match physical reality more closely with fewer falsehoods than previously believed.

Figure 9 maps the areas of greatest difference between the major cities gazetteer and the georeferenced baseline by dividing the world into a 1x1 degree grid and computing for each cell both the percent of all georeferenced tweets and the percent of all geocoded tweets found in that cell. The difference between the two percentages is calculated for each cell, indicating areas where the geocoding results deviate the most from the georeferenced baseline. Negative values (blue) indicate areas with fewer geocoded tweets than expected given the baseline and red indicates areas with more geocoded values than expected. Only differences greater than 0.05 percent are displayed. Overall the major contributors to the deviation between the geocoding and georeferenced results are the eastern U.S. and western Europe, where there are fewer geocoding coordinates compared with what georeferenced tweets would suggest. Central America and Indonesia both have an overage of geocoded tweets compared with georeferenced ones. India and the continent of Africa have relatively few differences, with those being more geocoding results than the baseline. The two areas of greatest deviation in Europe are the U.K. and southern Spain.



Figure 9: Percent difference between georeferenced and geocoded Twitter Decal 2012 to 30 November 2012. To see a higher resolution version of this figure, go [/go/twitter/images/hires/figure9.png](#).

Figure 10 shows there is massive variation by hour in the percentage of tweets matched by the geocoding system. From a peak of 68.9 percent of all tweets geocoded at 1AM PST to a low of 15.9 percent of tweets at 7PM PST, the textual geographic density of Twitter changes by more than 53 percent over the course of each day. This has enormous ramifications for the use of Twitter as a global monitoring system, as it suggests that the representativeness of geographic tweets changes considerably depending on time of day. The magnitude of this difference suggests a key driving factor may be the availability of English tweets in each region. Yet, as Figure 8 demonstrated, English tweets appear to match the broader geographic distribution of tweets in all other languages. A manual examination of a random sample of tweets processed by the geocoder reveals a surprising alternative explanation: just 43.5 percent of tweets with locations recovered by the geocoder are in English, with the remainder coming from non-English tweets that have English-language Location fields. For example, the Spanish spelling of Seoul, South Korea is "Seul, Corea del sur" yet the majority of Spanish tweets use the English spelling in their Location field. Table 3 shows the percentage of tweets in each language processed by the geocoder.

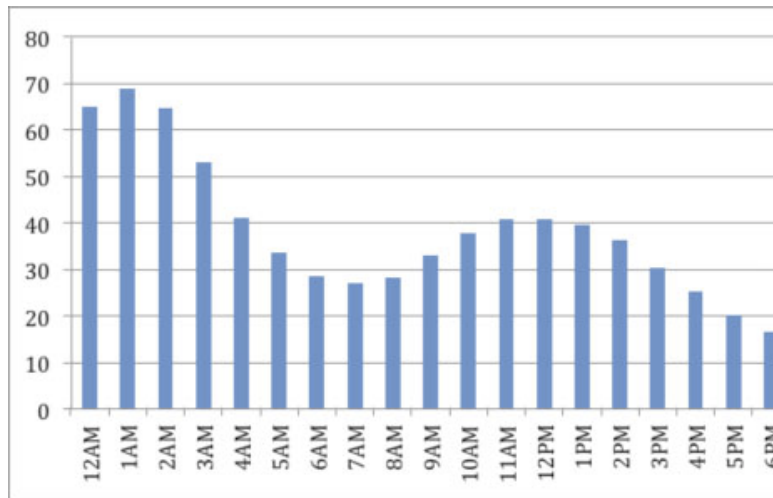


Figure 10: Percent of Twitter Decahose tweets geocoded by hour 23 October 2012 (PST).

	Percentage Tweets in this language with Location field	Percentage Tweets in this language with Location field geocoded
Greek	73.30	64.69
Spanish	69.08	63.50
French	71.39	60.62

Indonesian	79.25	59.74
Polish	72.12	59.47
Norwegian	70.97	58.42
Swedish	69.36	58.22
English	72.64	57.68
German	69.78	55.92
Finnish	68.17	55.48
Italian	73.39	51.97
Other	67.28	51.79
Turkish	63.15	51.07
Dutch	68.11	46.46
Portuguese	67.49	46.05
Arabic	65.74	41.21
Russian	51.36	38.66
Korean	63.32	18.16
Japanese	69.88	9.02

To explore this further, 100 tweets were selected at random from the two of the most common languages, Spanish and Indonesian, and their Location fields were passed through Google Translate and cross-checked with Web searches, local editions of Wikipedia, and multiple online translation dictionaries. For Spanish, 73 of the tweets had English-language Location fields, including both valid locations and phrases such as "somewhere", "all around the world", and "right in the middle." Of these, 59 had recoverable geographic locations, while two Spanish-language Location fields had recoverable locations. Indonesian tweets were nearly identical, with 74 Location fields in English, of which 59 had valid locations and an additional three Indonesian-language Location fields had recoverable locations. Japanese, with the lowest recovery rate of only nine percent, had just 11 English Location fields, of which nine contained recoverable locations. When translated from Japanese into English using Google Translate, 65 of the tweets contained recoverable locations. This would at first suggest that perhaps languages with Latin character sets might yield higher match rates than those of other character sets. To test this, a random selection of 100 tweets from Arabic were tested in similar fashion, with 65 Location fields in English, of which all 65 had recoverable locations, with an additional 15 Arabic-language Location fields containing recoverable locations. Greek, with its own character set, actually has the highest match rate of any language, and Russian, which uses the Cyrillic character set, also has an extremely high match rate of 39 percent. Thus, while languages using the Latin character set do appear to have slightly higher match rates, this does not appear to be the primary driving force.

Instead, it would appear that across the languages of the world, people carry on conversations in their native languages, but more often than not provide their location in English. An examination of the User Profile field shows that it overwhelmingly matches the language of the tweet itself, with just 25 percent of Arabic profiles containing any English, for example. The Location field often contains subtle unique additional text such as "Paris — The City of Light!" or "Tokyo my home of homes," indicating that the field is not being set to English through a software menu, but rather through users actually manually entering English text. Thus, it would appear that in most languages, users communicate with others speaking their language, describe themselves via their profile to others speaking their language, yet offer their location in English. One possible explanation is the recognition that software algorithms are increasingly attempting to map tweets through English-language processing of key metadata fields, especially the Location field, and so this represents an attempt by users across the world to ensure their tweets at least appear in search results. In a way, this would suggest that users don't care that English speakers know what they are saying, but they do want them to know they exist.

Asia appears to be the sole outlier, with the majority of Japanese and

Korean tweet Location fields containing valid locations, but only a few containing English-language locations. The two languages are also second and third last, respectively, in terms of the percentage of their tweets that are natively georeferenced, so it is not that geographic information is being provided through other means, but rather than users tweeting in those languages simply do not provide their locations in a form readily accessible to outsiders. Whether this reflects different cultural norms, a desire to focus on a more domestic audience with their tweets, or other reasons, is unclear. However, one possibility is that "Japanese internet users' preference for anonymity" and the strong presence of competing domestic social media platforms may reflect less of a desire for international visibility of user location (Elliott, *et al.*, 2012).

Comparing the percent of all tweets published in each language from [Table 2](#) with the percent of those tweets with recoverable English-language geographic information from [Table 3](#), it becomes clear that language plays a key role in the geocoder's coverage. The fact that the geocoder is able to recognize geographic locations in just a third of all tweets is therefore primarily due to the low levels of English-language Location fields in some of the more prevalent languages on Twitter. For example, Japanese comprises nearly 12 percent of all tweets, yet just nine percent of Japanese tweets contain an English-language Location field with recoverable geographic information. This is also the reason that the percentage of tweets the geocoder can process varies so dramatically through the course of the day, with the lowest point, 7PM PST, corresponding with noon in Japan.

Yet, one potentially bright spot is that since so much locative information is present in the Location field, which can be processed using a simple phrase-match into a city database, rather than requiring a full grammatical translation, it may be possible to vastly expand this match rate by incorporating the native language spellings of those major cities. Indeed, the GNS database already contains the local spellings of each city, so this would require relatively minor changes to the geocoding process. Alternatively, after verifying the Google Translate results with online phrase dictionaries, not a single translation yielded a location that had not been properly translated by Google Translate. This suggests that machine translation has reached a point where it is accurate enough for the coarse geocoding task required to geolocate tweets. This means that even if more advanced translation is required for accurate geocoding of some languages, machine translation, with its ability to scale to the velocity and size of Twitter, is accurate enough. It also means that even English-only geocoding systems will still successfully process a significant number of non-English tweets. However, the emerging use of Twitter to monitor societal unrest and natural disasters (Fraustino, *et al.*, 2012) will face an uneven landscape where some areas of the world are more strongly represented than others.



The geography of communication on Twitter

A key question explored in the communications literature over the decades has been the extent to which physical proximity enables or constrains human interaction and collaboration. In the traditional realm of "face to face" communication, users may speak only with those physically near them. The Internet era on the other hand has created a world in which a person may speak to another on the other side of the planet with just a few millisecond delay, effectively removing the geographic barrier. Yet, even as technology has made it possible to more readily communicate over distance, there has been a rich body of literature exploring whether these technologies have actually altered human communicative behavior, or whether distance still plays a critical role in facilitating social bonds and enriching communication. Scholars have explored geographic distance in scientific collaboration (Kraut, *et al.*, 1988), its effect on conflict (Hinds and Mortensen, 2005), and how it affects persuasion and deception (Bradner and Mark, 2002), among the countless areas studied. However, the majority of this literature is based on small case studies in professional contexts such as academic or corporate settings that may not necessarily reflect real-life utilization. Twitter offers the unique opportunity to re-explore these questions through the actual day-to-day social

interactions of the general public on a global scale. At the same time, it offers the ability to examine whether these patterns could be used to expand the percentage of tweets that can be geocoded by exploiting communicative locality. For example, if users communicate most regularly with users nearest to them, tweets from users with unknown locations could be assigned a location based on the average of all of the users that person interacts with on a daily basis.

Using Twitter to explore communicative distance is not a new concept, with Takhteyev, *et al.* (2012) finding that users emphasize ties with others within their same metropolitan region. In their paper they used tweets from August 2009 and selected a small sample of just 1,953 pairs of users, manually reviewing the Location field for each and hand geocoding them. They found that 75 percent of users had recoverable geographic information. This is substantially higher than found even through a manual review of the October–November 2012 Decahose stream, so it is unclear whether fewer users are reporting locations in the three years that have elapsed since this paper was written, or whether this higher density was a result of Twitter being far more U.S.–centric in 2009. The authors use these geocoded pairs of users to explore patterns in “following” behavior in which a user on Twitter can choose to “follow” another user, in effect “subscribing” themselves to receive a stream of all new tweets by that user. Following is a unidirectional action: a user can choose to “follow” another user and thereby receive all of that user’s new tweets, but the user being followed is under no obligation to follow or otherwise pay attention to the new followee.

Tracing followers has been widely used in the Twitter literature as a proxy for message consumption, exploring who is paying attention to whom. Yet, “follower” behavior yields a more nuanced understanding of consumption and attention in that it is a one–time action: a user clicks a button to follow another user just once and there is no way to stratify followed users to see which ones a user pays attention to the most. For example, the follower graph will show only that a user follows a given set of users, meaning she sees all tweets from those users, but not whether she cares more about tweets from certain users she follows than others. Instead, a more precise measure of user engagement is the volume of active communication between users. A user who actively messages another user over time, even if those messages are unidirectional, conveys a far greater sense of engagement than a user who simply allows another user’s messages to appear in her daily message stream.

Twitter is based around the broadcast communicative model where users publish a public message to be read by all other users of the service. However, users can also carry on public conversations with other users by referencing each other by username to direct a comment or question to a specific user or group of users (though any user can access these messages). Users can also rebroadcast or forward a tweet and in the process note the username of the user who posted the tweet originally, called “retweeting”. Both are highly popular on Twitter, with 55.1 percent of all tweets mentioning another user and 23.9 percent of all tweets including a retweet notation. On a typical day in the Decahose, 8.9M unique users are referenced (12.4 percent of all users active during this period) and 3.2M unique users are mentioned in retweets. This suggests that Twitter contains substantial information on realtime conversations among users. Overlaying this conversation onto the geographic strata of Twitter allows quantitative measurement of whether users tend to reference or retweet users nearer or farther from them. Is it the case that users communicate primarily with users physically near them, or is there little geographic affinity on the service?

The complete corpus of 1.5 billion tweets was processed to compile a list of tweets sent by each of the 71.3 million users who tweeted during this period. Each user’s tweets were scanned in chronological order to find the first georeferenced tweet, which was then assigned as the user’s location. If the user did not send any georeferenced tweets, the first geocoded tweet was assigned as the location. Since a quarter of all users sent only a single tweet during this period and more than half sent less than four, it was decided not to attempt a more complicated method of averaging all locations from each user in the case that a user sent multiple tweets containing geographic information. This resulted in a total of 27,142,286 users for which either a georeferenced or a geocoded location was

assigned, representing 38 percent of all users in the Decahose that sent a tweet during the period of analysis.

As a form of active republication, retweets offer a measure of a user's influence in the online sphere: those whose tweets are retweeted often by other users are actively resonating with their audiences to the point of motivating them to share those thoughts with others. In all, there were a total of 279,516,957 unique retweet pairings of users, defined as one user who has retweeted another user one or more times. A user which retweets another user multiple times is considered a single "pairing" for the purposes of this analysis. Looking only at those retweet pairings where the location of both users is known, there were 32,458,865 pairings where both users had Exact Location information and 70,248,089 pairings in which both had geocoded locations, meaning that 37.4 percent of all retweet connections between users were between two users both of whose locations are known.

Looking only at the locations of users who had been retweeted at least once, [Table 4](#) lists the top 20 cities with the most retweeted users. The United States accounts for seven of the entries, illustrating the outsized influence of its users. If a retweet is a measure of the "influence" of the original message by virtue of other users taking action to share it, then this is a list of the most influential cities on Earth. Dividing the world once more into a 1x1 degree grid, [Figure 11](#) shows the geographic distribution of retweets. When correlated against the total number of tweets from each location, the two are related at $r=0.93$, showing that the more tweets originating in a given city, the more tweets from that city that are retweeted. This suggests that geography plays little role in the location of influential users, with the volume of retweets instead simply being a factor of the total population of tweets originating from that city.

City	Percentage georeferenced retweets
New York	4.57
São Paulo	3.05
London	2.98
Paris	2.61
Kuala Lumpur	2.49
Los Angeles	2.20
Chicago	1.78
Bangkok	1.41
Jakarta	1.37
Rio de Janeiro	1.34
Dallas	1.30
Philadelphia	1.28
Madrid	1.23
Singapore	1.21
Houston	1.19
Riyadh	1.02
Istanbul	1.01
Porto Alegre	0.93
Toronto	0.72
San Antonio	0.70



Figure 11: Most retweeted cities (georeferenced Twitter Decahose tweets 23 November 2012). To see a higher resolution version of this figure, go to <http://www.firstmonday.org/ojs/index.php/fm/article/view/42111> /images/hires/figure11.png.

Turning now to the distance between the sender of the original tweet and the user who retweeted it, the Great Circle Distance, which takes into consideration the curvature of the Earth's surface, is calculated between each pair of users. The average distance between all 32.5 million retweet pairings in which both users have known Exact Location positions is 749 statute miles. Of those pairings, 80.7 percent have just a single communication, 11.2 percent communicate twice, and 3.7 percent communicate three times. When switching from Exact Location pairings to geocoded pairings, the average distance is 1,115 statute miles, reflecting the lower geographic resolution of the city centroids upon which it is based.

As opposed to retweets, referencing another user in a tweet represents a different type of bond, a form of public conversation, rather than a republication of a specific thought. There were a total of 588,505,838 unique reference pairings of users in the dataset, defined as one user referencing another user in the form "@username." As with retweets, a user who references another user multiple times is considered a single pairing for the purposes of this analysis. Of these, 50,860,630 were between users who both had Exact Locations and 127,236,509 were between users where both had geocoded locations. In all, 30.3 percent of all referencing pairings were between users where both had available geographic information. The average Great Circle Distance between users with Exact Locations was 744 statute miles, while the distance between geocoded users was 1,118 miles.

These distances average all pairings of users, whether they were connected by a single tweet or a hundred tweets. Instead, users who are physically proximate might communicate more often than users who are far away. [Figure 12](#) tests this theory by charting the average distance in statute miles between a pair of georeferenced users by the number of tweets connecting them. The X axis is the number of times the two users were connected, while the Y axis is the average distance in miles between them. A user who retweeted or referenced another user just a single time would count under X=1, while a pair of users connected 20 times would be tallied under X=20.

Users who retweet or reference another user just once are seen to have an average distance of around 800 miles, which decreases exponentially with the number of connections through a minimum of 605 miles for retweets and 524 miles for references. These minimums occur for users who are

connected nine times, with distance increasing once again for users who are connected more than nine times. There were less than 100 users each for $X > 61$ for retweets and $X > 116$ for references, but it can be seen that the average distance for reference pairings continues to increase through these maximums. The results are identical for geocoded users. The fact that the average distance between a pair of users decreases the more often they communicate strongly supports the notion that users communicate more often with those closest to them. Yet, the fact that this distance then increases linearly after nine connections suggests that users who communicate more than this are more likely to be connected with celebrities, public figures, and others for whom distance is less important.

Even at their closest, a 600 mile distance between retweeting users is still farther than the distance from San Diego to San Francisco, or driving across France from its northernmost to southernmost points. In Europe, 600 miles would connect users living in different countries. Thus, assigning a user with no available location information to the location of the people that user communicates most often with would likely result in significant error, potentially even assignment to the wrong country.

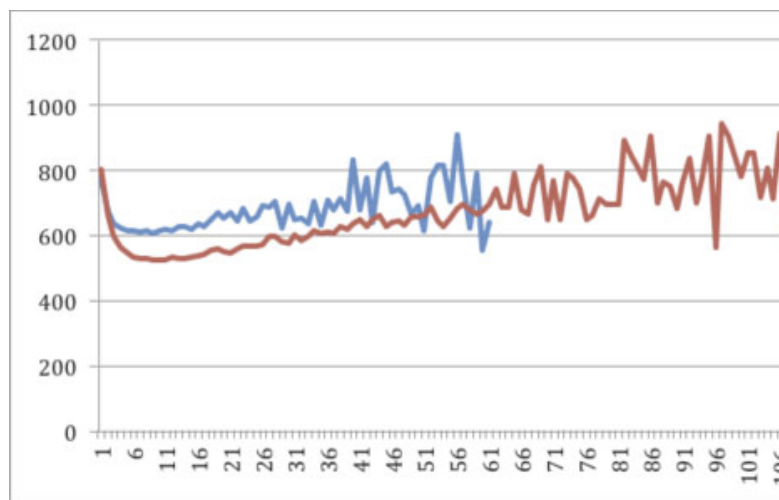


Figure 12: Average distance in miles between users by number of tweets connected and references (georeferenced Twitter Decahose tweets 23 October 2012 to 3

Average communicative distance offers a quick measure of how the number of connections between a pair of users relates to their physical proximity. However, it leaves unanswered the question of what the actual geographic network of interconnected users looks like. Is the minimum average distance of 600 miles between retweeting users due to all users having a distance around 600 miles, or is it because half of users have much shorter distances and the other half have far longer distances? [Figure 13](#) and [Figure 14](#) explore this by dividing the world into a 1x1 degree grid and aggregating all pairings of users by grid cell, to reduce the total number of connections to be drawn. Figure 13 focuses on georeferenced users, yielding a total of 1,004,955 connections among grid cells. This this would result in too many lines to display, so only connections with 100 or more pairings of users across the two cells were kept, which results in 42,650 connections in the final image. Since this map relies on natively georeferenced users, it captures connections across users tweeting in any language and thus maps cross-lingual connections. The map immediately makes clear that users do indeed retweet users near them, so there is substantial communicative locality. Yet, also immediately clear is that users also frequently retweet users far away from them, often on other continents. Latin America is more closely connected to Europe than to the United States, while Asia connects more closely to the U.S. and the Middle East connects to both the U.S. and Europe. The east coast of the United States is a clear nexus point for the country, through Europe

appears to be more dominant than the United States in producing content retweeted by the rest of the world.

Figure 14 repeats this process, but uses pairings of geocoded users instead of georeferenced users. The geocoder's reliance on city centroids results in a smaller number of connections between cells, 653,470 total links, which was reduced to 46,873 edges when only cells with more than 100 pairs of users were kept. This map exhibits even stronger inter-continental linking, explaining the higher average distance. Repeating this analysis for references, rather than retweets, yields a nearly identical image and the two are correlated at $r=0.98$, indicating that both retweet and reference pairings exhibit nearly identical linking patterns.

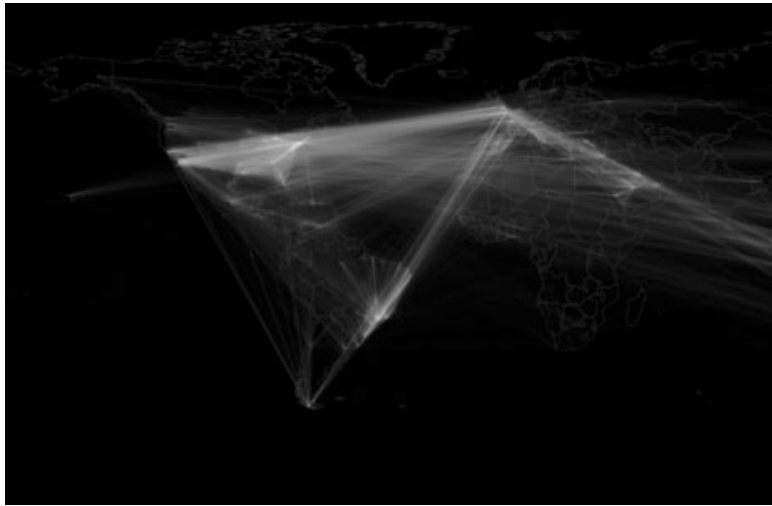


Figure 13: Network map showing locations of users retweeting other users (georeferenced users) (geocoded tweets 23 October 2012 to 30 November 2012). To see a higher resolution version, go to <http://www.sqi.com/go/twitter/images/hires/figure13.png>



Figure 14: Network map showing locations of users retweeting other users (geocoded users) (geocoded tweets 23 October 2012 to 30 November 2012). To see a higher resolution version, go to <http://www.sqi.com/go/twitter/images/hires/figure14.png>

On a typical day, 23.4 percent of all users who send a tweet or are referenced in a tweet, where no geographic information is available for that user, are paired in a reference pairing with a user whose location is known. This suggests that users without location information could be assigned the average location of users who retweet or reference or who are retweeted or referenced by that user. However, as the results in this section have illustrated, even at best this would result in an average error of more than 600 miles and would have a worst-case error of half the radius of Earth. From the network diagrams above, it is clear why: users actively communicate with those near them, but also actively communicate with those far from them, including on other continents. It could be possible to use additional criteria, such as leveraging the fact that three quarters of tweets include timezone information to only assign location to an unknown user based on users he or she is connected to within the same timezone. Yet, even this would not be sufficient, as the maps above show that South Americans connect to those in North America. Thus, communicative affinity is simply not strong enough to enable communicative networks to be used as a proxy for geographic location.



The geography of linking discourse

Users on Twitter communicate with others both near them and half a world away, illustrating that the role of physical proximity in communication seems to be reduced in the era of social media. Given that users communicate frequently both with those around them and those across the world, this raises the question of what they talk about. Does geographic affinity still exist in terms of the topics that users discuss, with users tweeting about events and activities closer to them? This would suggest that geography still plays an important role in that users are still local citizens with local events affecting them. If, on the other hand, users communicate just as frequently about events far away than near, it would argue for a far more globalized view of society in which geographic boundaries now play a much lesser role.

Given the 140-character limit of tweets and the findings of the geocoding experiments that the majority of tweets do not contain substantial geographic information, how does one determine the geographic focus of a tweet? Users may mention a location explicitly by name in a tweet (such as "horrible news about New York City and Hurricane Sandy"), but more often tweets simply encode non-referential emotional content ("what a terrible day") such that there is no clear indicator as to what is driving the concern. The concern may also be spread over multiple tweets: one tweet might say "praying New York City gets through Hurricane Sandy," while subsequent tweets might say "what terrible damage" and "hoping for a speedy recovery". One possibility is to exploit the fact that 15.85 percent of all tweets contain hyperlinks to external Web pages, which offer a far larger pool of text to mine for geographic information. If users tend to share web pages discussing locations near to them, this would be a strong indicator that geographic affinity plays a role in what they talk about, even if it plays less of a role in who they talk to. In this model, one could imagine Twitter's users as a mobile sensor network in which users talk to users across the world to access local information around where each user is based.

Hyperlinks in tweets are normally encoded using "URL shorteners" like bit.ly, which produce a shortened version of the URL to conserve space. Unfortunately, each time a URL is shortened, it is given a different unique identifier, meaning that multiple tweets referencing the same URL will appear to link to different URLs. The GNIP Decahose stream therefore automatically "resolves" all shortened links, converting them back to their original URL. There is no difference in the percent of tweets containing a hyperlink between georeferenced, geocoded, and users without geographic information, or between weekday and weekend tweets. However, there is a substantial difference in the density of links by time of day, ranging from a low of 12.7 percent at 7PM PST through a peak of 18.6 percent at 2AM PST. This is likely driven primarily by the area of the world using Twitter at any given moment.

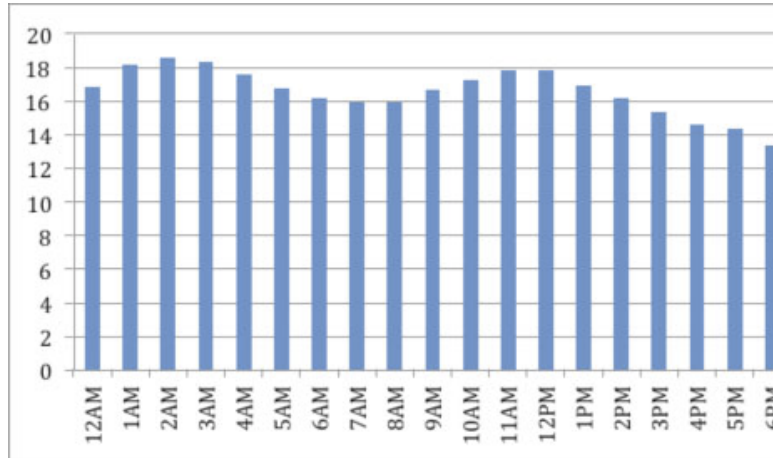


Figure 15: Percent all georeferenced tweets containing a link by hour (Twitter October 2012 to 30 November 2012) (PST).

In all, there were 485,941,182 links to 223,712,255 distinct URLs from 4,816,802 different Web sites (tweets can contain multiple links). The top six domains with the most links are twitter.com (16.8 percent), instagram.com (13.3 percent), facebook.com (11.9 percent), youtube.com (6.2 percent), ask.fm (3.2 percent), and tumblr.co (2.9 percent). FourSquare is number 8 with 2.5% and Flickr is number 65 at 0.1 percent. Looking just at georeferenced tweets, there were a total of 8,943,092 links to 7,331,672 distinct URLs from 113,389 Web sites. The top domains were foursquare.com (45.5 percent), instagram.com (17.5 percent), twitter.com (15.3 percent), myloc.me (3.5 percent), path.com (2.2 percent), and youtube.com (1.8 percent). Twitpic.com is number 10 at 0.4 percent, facebook.com is number 11 at 0.4 percent, and flickr.com is number 19 at 0.2 percent. Geocoded tweets contained 126,303,179 links to 67,135,720 distinct URLs from 2,071,802 Web sites. The top domains were twitter.com (13.8 percent), instagram.com (13.2 percent), facebook.com (12.5 percent), youtube.com (6.2 percent), foursquare.com (3.3 percent), and tumblr.co (2.2 percent). Twitpic.com is number 13 at 1.1 percent and flickr.com is number 54 with 0.1 percent. It is clear that multimedia and location-based services dominate tweeted links on Twitter for accounts with known geographic information, as well as links to other social media. In particular, the fact that nearly half of all links shared by users with geolocation enabled are to foursquare.com suggests those users leverage broadcasting their location as part of their general communicative stream. That links to other tweets and Twitter user profiles are so popular (accounting for nearly 17 percent of non-geographic tweets) is intriguing, as it indicates users are directly linking to statuses instead of using the @USER referencing conventions. This suggests that analyses of Twitter that rely purely on @USER tags will not catch all inter-user referencing. It is also quite interesting that Facebook links account for just 0.4 percent of all georeferenced tweets, yet account for 12.5 percent of geocoded tweets. The high percentage of links in georeferenced tweets to foursquare.com indicates why a third of those tweets did not have a language assignment from GNIP, since these tweets contain only a hyperlink with no surrounding text to allow language identification.

The fact that the landscape of the most popular tweeted Web sites emphasizes multimedia content and social media links like Facebook, presents a unique challenge to extracting geographic information. Multimedia usually has little text other than a small caption and keyword list, while social media links, such as Facebook profiles, are often password-protected. In contrast, English-language online news articles from Google News are rich in recoverable geographic information and well-suited for automated geocoding (Leetaru, 2011). Thus, as a comparison sample, all URLs listed in the RSS streams of the Google News

front page, primary topical pages, and individual country pages (via the "location:" feature) were collected during the same 23 October 2012 to 30 November 2012 period as the Twitter data. This yielded a list of 17,373 fully qualified Web site domains from which English-language Google News articles originated during this period. Links to URLs from these domains can therefore be assumed to be links to news coverage. Rescanning the list of all tweeted URLs for links from these domains, there were 17,416,210 links to 4,439,652 distinct URLs from 16,210 of the Web sites. Restricting to georeferenced tweets, there were 113,621 links to 58,396 distinct URLs from 4,642 distinct news Web sites, while for geocoded tweets there were 6,987,480 links to 2,384,559 distinct URLs from 15,040 distinct news sites.

Mainstream English-language news constitutes an extremely small portion of links, accounting for just 7.8 percent of all links, 3.6 percent of links from geocoded tweets, and 0.8 percent of georeferenced tweets. Unlike the broader set of all domains, georeferenced tweets, geocoded tweets, and non-geographic tweets all have the same four top news domains: bbc.co.uk, huffingtonpost.com, nytimes.com, and guardian.co.uk, making these the most popular English-language news Web sites on Twitter. To explore how geographic proximity affects the news that users share on Twitter, all georeferenced tweets were filtered for links to English-language news articles (defined as a URL from a domain which Google News included at least one link from during this period), resulting in a total of 18,650 URLs that were downloaded. The body text of each was extracted from the surrounding navigation bars and advertisements and passed through the fulltext geocoder engine. All locations extracted from each page were processed to find the one closest to the user's location using the Great Circle Distance and saved as its minimum distance.

A manual review of a random sampling of the URLs showed that many were national or international news stories of broad global interest. For example, a user in Chelsea, United Kingdom, tweeted a link to an article in the *Times of India* regarding a hanging in New Delhi, India, (*Times of India*, 2012), while a user in Bangalore, India, tweeted a link to a *The Verge* story on a new TechCrunch service that tracks Washington, D.C. policy (Jeffries, 2012). A user in Murmansk Oblast, Russia linked to a German article about Antarctica (Seidler, 2012), while a user in Ottawa, Canada, tweeted an Australian article about research in China (Payne, 2012). Perhaps the most complicated geographic chain was a user near Dusseldorf, Germany who tweeted a link to a United Kingdom newspaper's story about an article in the China's *People's Daily* about a story in the United States' *The Onion* about North Korea (Associated Press, 2012). One of the most interesting aspects of these examples is that it illustrates that users appear to be agnostic about the physical location of the news source they tweet, focusing on the story rather than the location of the outlet it comes from. This is in following with the continued transformation of newspapers from holistic information streams read from front to back to simple containers of stories (Leetaru, 2009).

The average minimum distance between user and the geographic focus of the article across all 18,650 news stories was 1,151 statute miles, in keeping with the large average distances seen in retweet and reference pairings among users in the previous section. Examining the distances more closely, just over a quarter of all links (26 percent) were to stories about the same city the user was located in, 37 percent were to events within a 100 mile radius of the user, and 47 percent were within 300 miles. At the same time, a nearly identical proportion (46 percent) were to stories about events more than 600 miles away, meaning that tweeted news stories were nearly evenly split between events near the user and those far away. This indicates that not only do users not preference communicating with users physically near them from those far away, but they discuss nearby and distant events at equal levels as well. This suggests that geography may play an even lesser role in social media than previously thought. It also shows that, similar to the previous section, users with unknown locations cannot be geolocated by assigning them the closest locations mentioned in news articles they link to.



User profile links

Another source of external links on Twitter comes in the form of Profiles, a free text field where users may include hyperlinks to other information about themselves. With just over 87 percent of all tweets posted from users with Profiles, this is a potentially rich source of additional linking information. In all, 3.3 percent of all tweets include a link in the Profile field and in just a 48-hour period, users linked to over 175,000 different Web sites from their profiles. Facebook was by far the most popular linked site, accounting for 30.5 percent of all profile links, followed by ask.fm (11.6 percent), twpf.jp (6.2 percent), and youtube.com (6.65 percent). VKontakte (a Russian equivalent to Facebook) was the tenth most linked site, with 1.1 percent of all profile links. The primary use of profile links thus appears to lie in connecting the user's Twitter account to his or her other social media accounts, some of which have their own geographic fields. This information could be leveraged to provide geolocation information for users that do not provide geographic information in their Twitter account by following these links and associating any geographic information from the link back to the Twitter account. Just under a third, or 30.5 percent, of all tweets that have links in their user profiles already have recoverable geographic information, thus, based on 3.3 percent of all tweets having profile links, this could translate to as much as a 2.3 percent gain in the number of geocoded tweets if every single profile link led to a Web page that included geographic information about the user.

Given that Facebook is the most popular linked site, a random selection of 20 linked profiles was manually reviewed, showing that nine were restricted with no details available to non-friends, but that the other 11 included the user's location in a publically viewable field. This would suggest that more than half of Twitter users who link to their Facebook profile include their location information on their Facebook persona. However, 39 percent of the tweets that link to Facebook profiles already have recoverable geographic information, and Facebook does not offer the same fine-grained remote user access API that Twitter does, and prohibits extracting profile information (Warden, 2010), making it difficult to recover this location information as part of a production environment. Thus, crawling links found in user profiles and associating that location information back to the original user account could increase the number of users with geographic information, but only by a very minor amount, and may not be possible based on the terms of use of many of these sites.

Another possible source of locative information would be to take the opposite approach and search the open Web for all mentions of Twitter handles and to triangulate the location of each user based on the most common locations discussed on Web pages mentioning that Twitter handle. However, this would be an enormous undertaking and feasible only for the largest Web indexes like Google. It could also yield highly misleading results, since a reporter might list his Twitter handle on every article he writes, associating his username with a vast array of locations across the globe.



Twitter versus mainstream news media

Another key question revolves around how Twitter's geographic profile compares with the mainstream news media it is often compared to. Does Twitter cover the same locations as the mainstream media, or do they discuss very different areas of the world? To explore this question, all 3,369,388 news articles monitored through the Google News RSS feed 23 October to 30 November 2012 were downloaded and subjected to fulltext geocoding. In all, 164,594 distinct locations were extracted, seen in [Figure 16](#). At first glance it is clear that there is some overlap with the earlier maps of Twitter's English geography, but that English-language mainstream media has a significantly different spatial profile that is far less dense and more diffuse. Gridding the world into a 1x1 degree grid, mainstream media coverage was correlated with the georeferenced Twitter baseline at just $r=0.26$, though with a 5x5 grid it was correlated at $r=0.67$, and with a 10x10 grid it was correlated at $r=0.74$.



Figure 16: All locations mentioned in articles found in the English-language Google Scholar database from October 2012 to 30 November 2012. To see a higher resolution version of this figure, go to <http://www.sgi.com/go/twitter/images/hires/figure16.png>

[Figure 17](#) overlays all mainstream news media locations (in red) on top of all georeferenced tweets (in blue). Areas that are blue have stronger Twitter representation, while red areas are covered more closely by mainstream media, and white areas have an equal balance. (This color scale is reversed from [Figure 5](#) because of the greater imbalance between the two datasets, in order to make the visual separation more clear.) Mainstream media appears to have significantly less coverage of Latin America and vastly better coverage of Africa. It also covers China and Iran much more strongly, given their bans on Twitter, as well as having enhanced coverage of India and the Western half of the United States. Overall, mainstream media appears to have more even coverage, with less clustering around major cities.



Figure 17: Comparison of georeferenced Twitter Decahose (blue) and English geographic coverage from October 2012 to 30 November 2012. To see a higher resolution version of this figure, go to <http://www.sgi.com/go/twitter/images/hires/figure17.png>

Following in the footsteps of Culturomics 2.0 (Leetaru, 2011), which explored mainstream news media coverage of the Arab Spring, especially Egypt, a comparison was made of the English-language discourse around Egypt in both Twitter and the mainstream news media captured by Google News. Textual mentions were compared first, capturing the percent of all tweets or news articles mentioning "egypt" or "cairo" each day. This results in a correlation of $r=0.77$, showing textual mentions of Egypt are highly correlated. However, as a second test, the total volume of geographic mentions was examined, comparing the percent of all tweets originating from users in Egypt on any topic each day with the percent of English Google News articles mentioning a location in Egypt that day. Comparing georeferenced tweets against the mainstream news yields a correlation of $r=0.26$, while using geocoded tweets yields a correlation of $r=0.48$. Time-shifting the two plots, to test whether the lower correlation is due to Twitter volume leading mainstream media, does not lead to an increase in the correlation.

These results indicate there is a strong difference in the geographic profiles of Twitter and mainstream media and that the intensity of discourse mentioning a country does not necessarily match the intensity of discourse emanating from that country in social media. It also suggests that Twitter is not simply a mirror of mainstream media, but rather has a distinct geographic profile, and that the differences between social and mainstream media geographic coverage deserve further exploration.



Twitter's geography of growth and impact

Two final questions revolve around the geography of Twitter's growth and influence: where is the service experiencing the greatest growth and which regions have the greatest influence on its discourse? Understanding where Twitter is growing is particularly crucial to exploring how quickly it is spreading to new areas of the world and thus its trajectory for reaching currently underrepresented locations. Similarly, simply knowing where the most tweets are posted from doesn't yield insight into influence: simply because a given city accounts for a large portion of tweets doesn't mean that anyone actually reads or engages with those tweets.

[Figure 18](#) divides the world in a 1x1 degree grid and calculates the average year that georeferenced users in that grid cell joined Twitter using the Twitter-provided user information metadata field. Grid cells with less than 1,000 tweets over this period were excluded to remove sparse areas. No cell had the majority of its users joining in 2012, indicating that even with Twitter's considerable growth it is not expanding that rapidly into areas it has not previously been used in. Immediately it becomes clear that the areas with predominately new users (growth areas) are in the Middle East, Western and Eastern Europe, Russia, and Asia. The western half of the United States, excluding the Pacific Coast, is the only area in the U.S. to have a significant proportion of new users. Latin America appears to be on par with the United States, while Europe has more areas of new users. India has one of the larger concentrations of areas with the longest Twitter users. Since this map is based on georeferenced users, it takes into account users of all languages.



Figure 18: Average year user joined Twitter (all georeferenced users posting a tweet in the Decahose 23 October 2012 to 30 November 2012). To see a higher resolution version of this map, visit <http://www.sqi.com/go/twitter/images/hires/figure18.png>

Yet, how influential are each of these regions? One measure is the average number of followers of Twitter users from a given area, seen in [Figure 19](#), again based on georeferenced users. This estimates the potential audience of each region, much as newspaper circulation offers a proxy for the potential number of people who could be reached by its coverage. In the United States, the two major centers with the highest numbers of followers are Los Angeles and the northeast, including New York City and Washington, D.C., all areas with high numbers of celebrities and public figures.



Figure 19: Average followers (all georeferenced users posting a tweet in the Decahose 23 October 2012 to 30 November 2012). To see a higher resolution version of this map, visit <http://www.sqi.com/go/twitter/images/hires/figure19.png>

However, the number of followers a user has is not necessarily indicative of

that user's total influence in the online sphere. How quickly the user is gaining new followers, how often his or her tweets are retweeted, and the role he or she plays in the overall network structure of Twitter all offer key dimensions of influence. To address this, a number of "social ranking" algorithms have been developed that attempt to measure more precisely the "impact" a given user has based on the spread of and interaction with that user's content through the online sphere. The GNIP Decahose includes the Klout (<http://klout.com/how-it-works>) score of each user, which [Figure 20](#) visualizes in a 1x1 degree grid as an alternative measurement of the actual influence and impact of each user (higher numbers indicate stronger impact). Immediately, strong geographic trends become evident in this map. Indonesia and Malaysia are nearly exclusively populated by users with the highest Klout scores, while France, the United Kingdom, and Spain have the highest density of high-Klout users in Europe. The Eastern United States and Venezuela have the highest in the Americas, while South Africa and Nigeria dominate Africa. The areas with the lowest Klout scores appear to be in the Middle East, Eastern Europe, and India. Turkey, in particular, has the largest density of low-Klout scores. This geographic stratification is significant in that it means that users in these areas play an outsized role in driving the overall Twitter discourse and messages or themes surfacing from users in these areas could serve as early warning indicators of subjects that may spread more rapidly across Twitter.



Figure 20: Average "Klout" score (all georeferenced users posting a tweet in the October 2012 to 30 November 2012). To see a higher resolution version of <http://www.sgi.com/go/twitter/images/hires/figure20.png>




Conclusions

This study has explored a month of the Twitter Decahose, 10 percent of the global Twitter stream, consisting of over 1.5 billion tweets from more than 70 million users. Just over three percent of all tweets include native geolocation information, with two percent offering street address-level resolution in real-time. Georeferenced tweets are correlated at $r=0.79$ with the NASA City Lights imagery, meaning where there is electricity, there are tweets. Yet, one percent of all users accounted for 66 percent of georeferenced tweets, indicating they capture the activity of just a small fraction of Twitter's user base. Country boundaries largely circumscribe the languages used on Twitter, with most countries having a dominant language other than English, though the distribution of English tweets is correlated at $r=0.75$ with all languages, indicating English offers a strong

geographic proxy of all discourse.

The small volume of georeferenced tweets can be dramatically enhanced by applying geocoding algorithms to the textual content and metadata of each tweet. Each metadata field available in the Twitter Decahose JSON data stream was tested for geographic information and a wide array of major geocoding algorithms and approaches explored along the dimensions of coverage and accuracy. The most accurate field is the self-reported user Location field, which yields a correlation of $r=0.72$ with the georeferenced baseline and increases the density of mappable tweets from three percent to 34 percent when combined with the Profile field. Even when users tweet in a language other than English, they tend to specify their Location in English, suggesting they don't care that English speakers know what they are saying, but they do want them to know they exist. Most critically, since the user Location and Profile fields do not change on a regular basis for most users, a geocoding system only has to geocode users, rather than tweets, meaning it would need to process just 70 million user accounts during the monitored month, rather than 1.5 billion tweets, and can store user location information in a lookup table.

There appears to be only weak geographic affinity in communicative link formation in that users retweet and reference users far away nearly as often as they do those physically proximate to them. Similarly, half of the news coverage tweeted by users was about events close to them, while a nearly equal amount was far away. This suggests that social media is indeed having a significant impact on the role of geographic proximity in the communicative landscape. Mainstream news media and Twitter are seen to have very different geographic profiles, but similar temporal patterns. Finally, the Middle East and Eastern Europe account for some of Twitter's largest new growth areas, while Indonesia, Western Europe, Africa, and Central America have high proportions of the world's most influential Twitter users. 

About the authors

Kalev Leetaru is a University Fellow at the University of Illinois Graduate School of Library and Information Science. His work focuses on "big data" analyses of space, time, and network structure to understand human society and culture through its communicative footprints and is the author of *Data mining methods for the content analyst: An introduction to the computational analysis of content* (New York: Routledge, 2012).
E-mail: kalev [dot] leetaru5 [at] gmail [dot] com

Shaowen Wang is an Associate Professor in the Department of Geography and Geographic Information Science, and founding Director of the CyberInfrastructure and Geospatial Information Laboratory at the University of Illinois at Urbana-Champaign.
E-mail: shaowen [at] illinois [dot] edu

Guofeng Cao is a Postdoctoral Research Associate in the Department of Geography and Geographic Information Science and the CyberInfrastructure and Geospatial Information Laboratory at the University of Illinois at Urbana-Champaign. His research interests include geographic information science, spatiotemporal analysis and location-based social media data analysis.
E-mail: guofeng [at] illinois [dot] edu

Anand Padmanabhan is a research scientist in the CyberInfrastructure and Geospatial Information Laboratory at the University of Illinois at Urbana-Champaign. His primary research interests include the development of large-scale distributed environments based on Grid computing, multi-agent systems, and P2P systems.
E-mail: apadmana [at] illinois [dot] edu

Eric Shook is a Ph.D. candidate in the Department of Geography and Geographic Information Science and the CyberInfrastructure and Geospatial Information Laboratory at the University of Illinois at Urbana-Champaign. His research interests include large scale agent-based modeling and high performance computing.
E-mail: eshook2@illinois.edu

Acknowledgements

The authors would like to acknowledge the support of Silicon Graphics International (SGI) for the use of one of their new UV2000 supercomputers and to GNIP for access to the Twitter Decahose. They would also like to acknowledge and thank the support of SGI staff David Berg, Thomas Borneman, John Kichury, Eugene Kremensky, Val Roux, Mike Travis, Bill Van Dyken, and Laura Wang.

References

- Associated Press, 2012. "China's *People's Daily* falls for Kim Jong-un 'sexiest man alive' spoof," *The Guardian* (27 November), at <http://www.guardian.co.uk/world/2012/nov/27/china-kim-jong-un>, accessed 16 January 2013.
- Mitja D. Back, Juliane M. Stopfer, Simine Vazire, Sam Gaddis, Stefan C. Schumke, Boris Egloff, and Samuel D. Gosling, 2010. "Facebook profiles reflect actual personality, not self-idealization," *Psychological Science*, volume XX(X), numbers 1–3, at <http://www.result.de/wp-content/uploads/2010/02/JGUM-Persoenlichkeit-2.0.enql.pdf>, accessed 16 January 2013.
- Erin Bradner and Gloria Mark, 2002. "Why distance matters: Effects on cooperation, persuasion, and deception," *CSCW '02: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, pp. 226–235, and at <http://www.ics.uci.edu/~gmark/CSCW2002.pdf>, accessed 22 April 2013.
- Carlos Castillo, Marcelo Mendoza, and Barbara Poblete, 2013. "Predicting information credibility in time-sensitive social media," *Internet Research*, at http://chato.cl/papers/castillo_mendoza_poblete_2012_predicting_credibility_twitter.pdf, accessed 16 January 2013.
- Paul Earle, Daniel Bowden, and Michelle Guy, 2011. "Twitter earthquake detection: Earthquake monitoring in a social world," *Annals of Geophysics*, volume 54, issue 6, pp. 708–715, and at <http://www.annalsofgeophysics.eu/index.php/annals/article/view/5364>, accessed 22 April 2013.
- Nate Elliott and Gina Sverdlov, with Reineke Reitsma, Amelia Martland, and Samantha Jaddou, 2012. "Global social media adoption in 2011 — A social computing report: A review Of Forrester's social technographics data from around the world," *Forrester Reports*, at <http://www.forrester.com/Global+Social+Media+Adoption+In+2011/fulltext/-/E-RES60605?docid=60605>, accessed 22 April 2013.
- Seth Fiegerman, 2012. "Twitter now has more than 200 million monthly active Users," *Mashable* (18 December), at <http://mashable.com/2012/12/18/twitter-200-million-active-users/>, accessed 16 January 2013.
- Eric Fischer, 2011. "Language communities of Twitter," at <http://www.flickr.com/photos/walkingsf/6277163176/>, accessed 16 January 2013.
- Julia Daisy Fraustino, Brooke Liu, and Yan Jin, 2012. "Social media use during disasters: A review of the knowledge base and gaps," *Final Report to Human Factors/Behavioral Sciences Division, Science and Technology Directorate, U.S. Department of Homeland Security*, at http://www.start.umd.edu/start/publications/START_SocialMediaUseduringDisasters_LitReview.pdf, accessed 16 January 2013.
- Brandon Griggs, 2012. "Twitter accounts for storm, relief updates," *CNN* (29 October), at <http://www.cnn.com/2012/10/29/tech/social-media/storm-sandy-social-media/>, accessed 16 January 2013.
- Pamela Hinds and Mark Mortensen, 2005. "Understanding conflict in geographically distributed teams: The moderating effects of shared identity, shared context, and spontaneous communication," *Organization Science*, volume 16, number 3, pp. 290–307. <http://dx.doi.org/10.1287/orsc.1050.0122>

- Adrianne Jeffries, 2012. "TechCrunch debuts CrunchGov, a guide to tech policy developed with Silicon Valley's help," *The Verge* (26 October), at <http://www.theverge.com/2012/10/26/3555960/techcrunch-launches-crunchgov-tech-policy>, accessed 16 January 2013.
- Yasmin Khorram, 2012. "As Sandy pounded NYC, fire department worker was a Twitter lifeline," *CNN* (1 November), at <http://www.cnn.com/2012/11/01/tech/social-media/twitter-fdny/>, accessed 16 January, 2013.
- Robert Kraut, Carmen Egido, and Jolene Galegher, 1988. "Patterns of contact and communication in scientific research collaboration," *CSCW '88: Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work*, pp. 1-12.
- Kalev Leetaru, 2012. "Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched Wikipedia," *D-Lib Magazine*, volume 18, numbers 9-10, at <http://www.dlib.org/dlib/september12/leetaru/09leetaru.html>, accessed 16 January 2013.
- Kalev Leetaru, 2011. "Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space," *First Monday*, volume 16, number 9, at <http://firstmonday.org/ojs/index.php/fm/article/view/3663/3040>, accessed 16 January 2013.
- Kalev Leetaru, 2009. "New media vs. old media: A portrait of the *Drudge Report* 2002-2008," *First Monday*, volume 14, number 7, at <http://firstmonday.org/ojs/index.php/fm/article/view/2500/2235>, accessed 16 January 2013.
- Library of Congress, 2013. "Update on the Twitter archive at the Library of Congress," at http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf, accessed 16 January 2013.
- MaxMind, 2013. "Free world cities database," at <http://www.maxmind.com/en/worldcities>, accessed 12 January 2013.
- Patrick Meier, 2012. "How the UN used social media in response to Typhoon Pablo," *iRevolution* (8 December), at <http://irevolution.net/2012/12/08/digital-response-typhoon-pablo/>, accessed 16 January 2013.
- Greg Miller, 2011. "Social scientists waded into the Tweet stream," *Science*, volume 333, number 6051 (30 September), pp. 1,814-1,815.
- NASA, 2000. "NASA visible Earth: Earth's city lights," at <http://visibleearth.nasa.gov/view.php?id=55167>, accessed 16 January 2013.
- Rob Payne, 2012. "Mathematicians suggest new way for aircraft boarding," *Science Network* (13 November), at <http://www.sciencewa.net.au/topics/social-science/item/1816-mathematicians-suggest-new-way-for-aircraft-boarding.html>, accessed 16 January 2013.
- Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes, 2011. "Do all birds tweet the same? Characterizing Twitter around the world," *CIKM 2011: 20th ACM Conference on Information and Knowledge Management* (Glasgow, Scotland), at <http://www.ruthygarca.com/papers/cikm2011.pdf>, accessed 16 January 2013.
- Christoph Seidler, 2012. "Norway's New Foreign Minister: 'Exploitation of Arctic Resources Will Happen'," *Spiegel Online* (26 October), at <http://www.spiegel.de/international/world/interview-norway-s-foreign-minister-espen-barth-eide-on-arctic-drilling-a-863558.html>, accessed 16 January 2013.
- SemioCast, 2012. "Twitter reaches half a billion accounts: More than 140 millions in the U.S." (30 July), at http://semioCast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US, accessed 16 January 2013.
- Gerry Shih, 2012. "Twitter and Nielsen pair up to publish new 'social TV' ratings," *Reuters* (17 December), at <http://www.nbcnews.com/technology/technolog/twitter-nielsen-pair-publish-new-social-tv-ratings-1C7660195>, accessed 16 January 2013.
- Brad Stone, 2012. "Twitter, the startup that wouldn't die," *Bloomberg*

- BusinessWeek* (1 March), pp. 62–67, and at <http://www.businessweek.com/articles/2012-03-01/twitter-the-startup-that-wouldnt-die>, accessed 22 April 2013.
- Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman, 2012. "Geography of Twitter networks," *Social Networks*, volume 34, number 1, pp. 73–81, and at <http://takhteyev.org/papers/Takhteyev-Wellman-Gruzd-2010.pdf>, accessed 22 April 2013.
- Times of India*, 2012. "Ajmal Kasab hanged day after UN vote on penalty" (22 November), at <http://timesofindia.indiatimes.com/india/Ajmal-Kasab-hanged-day-after-UN-vote-on-penalty/articleshow/17317333.cms>, accessed 16 January, 2013.
- Twitter, 2013a. "Connecting advertisers to Twitter users around the world" (22 January), at <http://advertising.twitter.com/2013/01/international-advertisers.html>, accessed 16 January 2013.
- Twitter, 2013b. "Twitter Help Center: FAQs about Tweet location," at <http://support.twitter.com/articles/78525-faqs-about-tweet-location>, accessed 16 January 2013.
- Twitter, 2013c. "Twitter Help Center: Adding your location to a Tweet," at <http://support.twitter.com/articles/122236-how-to-tweet-with-your-location>, accessed 16 January 2013.
- Twitter, 2013d. "Twitter Help Center: Using the location feature on mobile devices," at <http://support.twitter.com/articles/118492-how-to-tweet-with-your-location-on-mobile-devices>, accessed 16 January 2013.
- Twitter, 2011. "One hundred million voices" (8 September), at <http://blog.twitter.com/2011/09/one-hundred-million-voices.html>, accessed 16 January 2013.
- Twitter, 2009. "Twitter Blog: Location, location, location" (20 August), at <http://blog.twitter.com/2009/08/location-location-location.html>, accessed 16 January 2013.
- Pete Warden, 2010. "How I got sued by Facebook," *PeteSearch* (5 April), at <http://petewarden.typepad.com/searchbrowser/2010/04/how-i-got-sued-by-facebook.html>, accessed 16 January 2013.

Editorial history

Received 9 February 2013; accepted 19 April 2013.

Copyright © 2013, *First Monday*.
Copyright © 2013, Kalev H. Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook.

Mapping the global Twitter heartbeat: The geography of Twitter
by Kalev H. Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan,
and Eric Shook.

First Monday, Volume 18, Number 5 - 6 May 2013
<http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654>
doi:10.5210/fm.v18i5.4366

A Great Cities Initiative of the University of Illinois at Chicago [University Library](#).

© *First Monday*, 1995-2015.