Traveling Trends: Social Butterflies or Frequent Fliers?

Emilio Ferrara* Onur Varol

Filippo Menczer Alessandro Flammini

Center for Complex Networks and Systems Research
School of Informatics and Computing, Indiana University, Bloomington, USA

ABSTRACT

Trending topics are the online conversations that grab collective attention on social media. They are continually changing and often reflect exogenous events that happen in the real world. Trends are localized in space and time as they are driven by activity in specific geographic areas that act as sources of traffic and information flow. Taken independently, trends and geography have been discussed in recent literature on online social media; although, so far, little has been done to characterize the relation between trends and geography. Here we investigate more than eleven thousand topics that trended on Twitter in 63 main US locations during a period of 50 days in 2013. This data allows us to study the origins and pathways of trends, how they compete for popularity at the local level to emerge as winners at the country level, and what dynamics underlie their production and consumption in different geographic areas. We identify two main classes of trending topics: those that surface locally, coinciding with three different geographic clusters (East coast, Midwest and Southwest); and those that emerge globally from several metropolitan areas, coinciding with the major air traffic hubs of the country. These hubs act as trendsetters, generating topics that eventually trend at the country level, and driving the conversation across the country. This poses an intriguing conjecture, drawing a parallel between the spread of information and diseases: Do trends travel faster by airplane than over the Internet?

Categories and Subject Descriptors

[**Human-centered computing**]: Collaborative and social computing—*Social media*; [**Information systems**]: World Wide Web—*Social networks*; [**Networks**]: Network types—*Social media networks*

*Corresponding author: ferrarae@indiana.edu

COSN'13, October 7–8, 2013, Boston, Massachusetts, USA.

ACM 978-1-4503-2084-9/13/10 ...$15.00.
http://dx.doi.org/10.1145/2512938.2512956.

Keywords

Social media; Twitter; trends; geography; mobility

1. INTRODUCTION

Social media and online social networks have been widely adopted as proxies to study complex social dynamics, such as the spread of information and opinions [11, 16, 25, 29, 53, 54] and the emergence of patterns of collective attention [5, 6, 26, 50]. Groundbreaking results emerged with the analysis of geographic metadata from social media, allowing for the study of human mobility patterns and social media demographics [20, 24, 31, 35, 45, 46, 8].

It has been suggested that social media may overcome the spatio-temporal limitations of traditional communication: technologically-mediated systems make it possible to ignore physical and geographic distances [12, 34]. This, however, does not imply that communication patterns on social media are not affected by physical distances and geographic borders [33, 36]. In this paper, we explicitly study the role played by geography in driving the main topics of discussion on Twitter: trending hashtags and phrases.

Trends represent interesting collective communication phenomena: they are user-generated, continually changing and mostly ungoverned (although orchestrated hijacking attempts have already been observed [9, 40, 41]). So far, trends have been studied as a proxy to detect exogenous real-world events discussed in social media, [1, 3, 17, 43], emerging topics, or news of interest for the online community [10, 27].

But trends are also strongly localized in space and time: the temporal and geographic dimensions play a crucial role to determine the success of a trend in terms of spreading and longevity. We argue that unveiling the spatio-temporal dynamics that drive trending conversations on social media is instrumental to many purposes: from designing successful advertising campaigns, to understanding virality and popularity that characterize some topics. In this paper we characterize the relation between trends and geography by tracking and analyzing trending topics on Twitter in 63 main locations of the United States and at the country level, for a period of 50 days in 2013.

Contributions and outline

Here we study the distribution, origins, and pathways of trends; the dynamics underlying trend production and consumption in different geographic areas; and the competition among trends to achieve global popularity. In the remainder of the paper we make the following contributions:

Table 1: The list of the 63 trend locations in the United States and the relative total number of trends (thousands) they generated in the period between April, 12th and the end of May 2013.

Albuquerque	6.7	Cincinnati	5.8	Greensboro	5.8	Long Beach	6.5	New Haven	5.6	Pittsburgh	5.8	San Francisco	5.7
Atlanta	5.1	Cleveland	5.4	Harrisburg	6.3	Los Angeles	5.2	New Orleans	6.2	Portland	6.4	San Jose	6.6
Austin	5.8	Colorado Springs	6.7	Honolulu	6.5	Louisville	5.9	New York	4.4	Providence	5.9	Seattle	5.9
Baltimore	5.8	Columbus	6.0	Houston	5.1	Memphis	6.5	Norfolk	6.0	Raleigh	5.3	St. Louis	5.7
Baton Rouge	6.5	Dallas-Ft. Worth	5.3	Indianapolis	5.9	Mesa	6.6	Oklahoma City	5.8	Richmond	6.2	Tallahassee	6.3
Birmingham	6.1	Denver	6.1	Jackson	6.8	Miami	5.5	Omaha	6.4	Sacramento	5.9	Tampa	5.6
Boston	5.0	Detroit	4.8	Jacksonville	6.0	Milwaukee	5.8	Orlando	5.8	Salt Lake City	6.4	Tucson	6.6
Charlotte	5.2	El Paso	6.5	Kansas City	5.7	Minneapolis	5.6	Philadelphia	5.1	San Antonio	5.8	Virginia Beach	6.8
Chicago	5.2	Fresno	6.6	Las Vegas	5.4	Nashville	6.0	Phoenix	5.9	San Diego	6.2	Washington	4.7

- In §2.2 we describe a procedure to build a directed and weighted temporal dependence network to infer the trendsetting and trend-following relationships among locations.
- In §3.1 we provide a statistical characterization of trends, describing how they are distributed in space and time.
- In §3.2 we highlight a locality effect in the trend sharing patterns: geographically close cities share similar trends. This effect of locality yields the emergence of three geographic clusters in the US, namely East coast, Midwest, and Southwest. But we also uncover a surprising fourth cluster, representing metropolitan areas spread across the country.
- The temporal dependence network is exploited to unveil the pathways that trends follow: in §3.3 we reconstruct and reveal the significant backbone of this network that carries the trends across the country.
- In §3.4 we describe two different dynamics that govern popularity of trends at the country level, one for cities in each local geographic area and one for metropolitan areas. We conclude highlighting that the major metropolitan areas shape the country trends significantly more than all other locations in the country.
- Finally, in §4 we propose an interpretation for the trendsetting role of major metropolitan areas, by noting their correspondence with air traffic hubs and conjecturing that trends travel through air passengers, just as infectious diseases.

A more extensive literature review can be found in §5.

2. EXPERIMENTAL SETUP

In this section we discuss the methodology we followed to generate a dataset of Twitter trends, and the derived temporal dependence network that allows us to unveil the dynamics of trend production and consumption.

2.1 Trends dataset

To build our dataset we monitored in real-time all trends appearing on Twitter for a period of 50 days, starting from April, 12th until the end of May 2013.

The Twitter homepage provides a trends box that contains the top 10 trending hashtags or phrases at any given moment, ranked according to their popularity. Oftentimes, a promoted trend is showed in 1st position — for our analysis we disregarded promoted trends since their popularity is artificially inflated by the advertisement.

Each Twitter user can monitor the trends at the *worldwide*, *country*, or *city* level. Twitter has identified 63 locations in the United States, displayed in Figure 4, for which it is possible to follow local trends. The full list of locations is reported in Table 1. It is worth noting that some areas are over-represented (for example the East coast and California), while some states (namely, North and South Dakota, Montana, Wyoming, Idaho, and Alaska) are not represented at all.[1]

We deployed a Web crawler to check at regular intervals of 10 minutes the trends of each of these 63 locations and, in addition, those at the country level. We ended up collecting 11,402 different trends overall: 4,513 hashtags and 6,889 phrases. Table 1 also reports how many trends have been observed in each location.

2.2 Trend pathway backbone network

To investigate where trends usually start and how they propagate from city to city, we built a temporal dependence network of the 63 locations of the United States represented in our dataset.

This network is directed and weighted: each node corresponds to one of the 63 cities, and the weight of an arc e_{ij} from node i to node j is increased every time location i exhibits a trend before location j. The weight of arc e_{ij} therefore represents the extent to which city i precedes city j in adopting a trend: the higher the weight, the more often location i sets the trends that location j will later adopt.

Due to the fact that the adopted dataset contains a large number of trending hashtags and phrases, the network obtained using the procedure described above is fully-connected. This makes the extraction of relevant connections hard, as each location is connected with all the others and only the weight of the connections vary.

To ease the analysis we applied to this network an edge filtering technique known as multiscale backbone extraction [48]. The goal of this procedure is to retain only those connections that are statistically significant, by removing all edges whose weight does not deviate sufficiently from a null model. The significance level of an edge is determined by a threshold parameter α. Lowering α progressively removes edges and eventually causes the disruption of the network. We tuned α to obtain the backbone network with the minimum number of edges that suffices to maintain all 63 nodes connected ($\alpha = 0.3$). The resulting multiscale backbone of the network is used for the analysis of pathways of trend diffusion, and to investigate trendsetting and trend-following dynamics (see §3.3).

3. RESULTS

The results of our analysis are discussed in this section: after a statistical description of trends, discussing how they are distributed in space and time (§3.1), we explore their geographic dimension, defining what areas of the country share

[1]This has to do with the fact that the activity on Twitter in those states is very low.

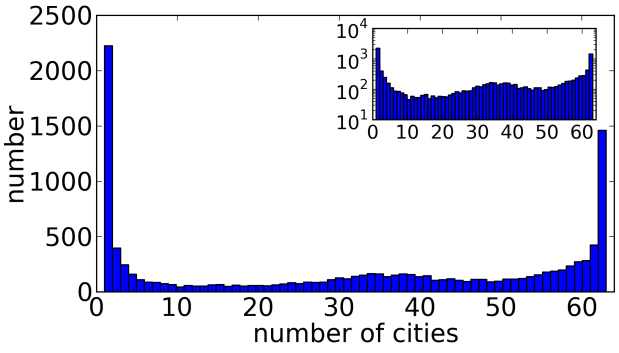

Figure 1: Histogram of the number of trends appearing in different number of places. Inset: y-axis reported in a log-scale.

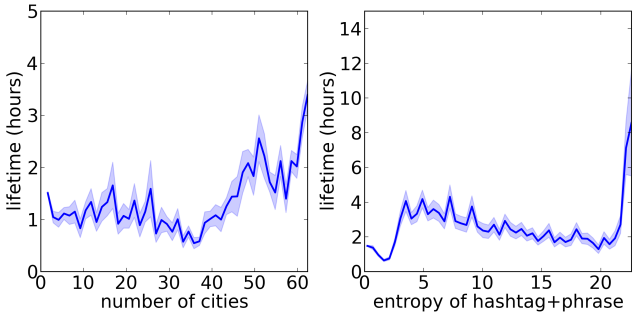

Figure 2: Lifetime of a trend. Left: as function of the number of cities in which a trend has appeared. Right: as function of its entropy. In both plots, the dark blue line is the average across trends while the standard error is depicted in light blue.

the same type of trends (§3.2); then we further investigate the temporal dimension, discussing the pathways trends follow (§3.3), and finally we characterize the trendsetting and trend-following dynamics (§3.4).

3.1 Spatio-temporal trend analysis

In our first experiment we aim to give a statistical characterization of trends: in particular, we start investigating in how many different cities trends appear. In Figure 1 we report the number of trends appearing in a given number of distinct locations. Trends follow a bimodal distribution, typically appearing either in one or few locations, or in all or most of them. We can identify three behaviors: (i) a large fraction of trends are localized and not sustained enough to spread from their originating place to others; (ii) another comparably large fraction of trends diffuse all over the cities generating a global phenomenon across the country; and (iii) the small remainder diffuse from the originating place to some other places, but fail to achieve global popularity.

The lifetime of trends is broadly distributed: short-lived topics trending for less than 20 minutes amount for more than 68% of the total, and overall trends shorter than six hours cover more than 95% of our sample. Sporadically some trends happen to live a much longer time, with only 0.3% surviving for more than a day.

We now focus on the spatio-temporal dimension of trends, aiming to determine how much time each trend spends in one or several locations. In particular, we calculate the average lifetime of a trend (the average amount of time a given hashtag or phrase is trending somewhere) as a function of the number of cities in which it appears. Figure 2 (left panel) reflects the intuition that trends reaching more places live longer.

Another way to determine the relation between the *geographic spread* of trends and their temporal patterns is to measure their lifetime as a function of *entropy*, defined as

$$\mathcal{S}^j = -\sum_i P_i^j \log P_i^j, \text{ with } P_i^j = \frac{t_i^j}{\sum_k t_k^j}, \quad (1)$$

where t_i^j is the time topic j has been trending in location i. The entropy is low if the trending topic is concentrated in a few places, and maximal if the topic trends for equal durations of time in all places. Figure 2 (right panel) shows that for trends with low entropy (*i.e.*, those concentrated in a single location), the expected lifetime is very short. The lifetime increases significantly (five-fold) for the maximum observed entropy. This analysis reveals a key ingredient for global trend popularity: the trending time of a topic is not only determined by its lifetime in a single location, but also by its geographic spread across many locations.

3.2 Geography of trends

Let us examine the geographic patterns of trends, namely whether geographically close cities share more similar trends than cities that are physically far apart. To determine if this locality effect exists, we first isolate, for each location i, the set of trends T_i that appeared in that location. Then, for each pair of locations i and j we compute the pairwise Jaccard similarity

$$S_{ij} = \frac{|T_i \bigcap T_j|}{|T_i \bigcup T_j|}. \quad (2)$$

The Jaccard similarity ranges between 0 and 1: the higher the value, the more similar the trends exhibited by two different cities. These values of similarity are subsequently passed to a hierarchical clustering algorithm after being transformed in distances: $d_{ij} = 1 - S_{ij}$. This is done to determine whether it is possible to isolate clusters of locations that exhibit similar trends, and, if so, whether these locations are geographically close or spread all over the country. The result is showed in Figure 3 and discussed next.

3.2.1 *Locality effects*

Figure 3 is constituted by two parts: a heat-map representing the pairwise Jaccard similarity among locations, and a dendrogram generated according to an agglomerative hierarchical clustering algorithm using complete linkage. Analyzing the dendrogram we can identify three distinct clusters, whose members (reported in different colors: green, yellow and red) share a high internal similarity in the trends exhibited during the observation period. This cluster emerges applying a cut to the dendrogram for a distance value of 0.5. We can also identify a fourth cluster (in purple, emerging with a dendrogram cut corresponding to a distance value of 0.75) that exhibits a lower internal similarity and whose members show a low similarity with those of other clusters. The four clusters are reported in Table 2, and displayed in Figure 4.

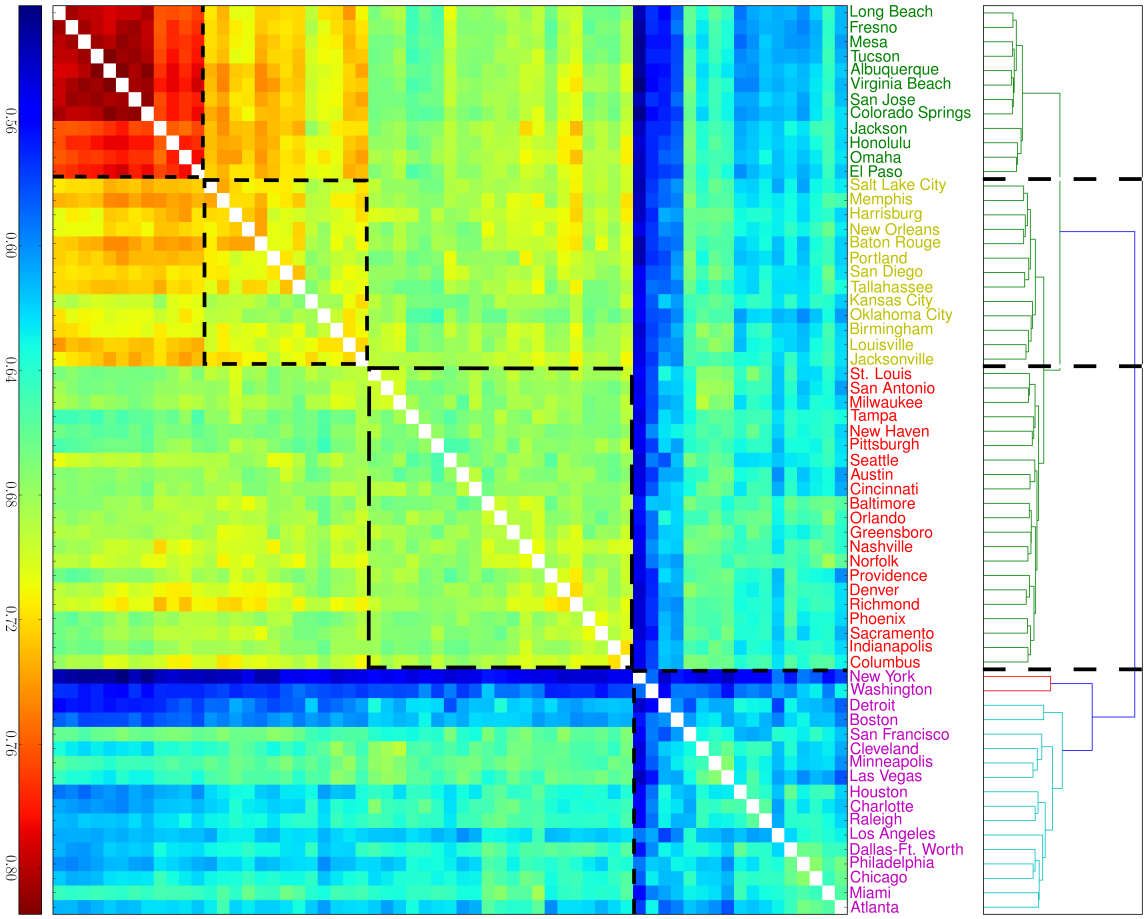

Figure 3: Shared trend similarity and hierarchical clustering of the 63 locations.

From the figure we observe that the green, yellow and red clusters are somewhat geographically localized, while the purple one is spread more or less all over the country. In detail, the green cluster, with the highest internal similarity, roughly corresponds to the Southwest of the country. The yellow cluster follows, representing the Midwest and South. The red cluster, which is less localized, matches many locations in the East coast and Midwest. The purple cluster includes several major metropolitan areas [51]; their effect on trendsetting dynamics is discussed in §3.4 and a conjecture about their role is offered in §4.

3.2.2 *Significance of geographic clustering*

To determine the statistical significance of the clustering obtained by using the previous method we proceeded as follows: we first computed the distribution of similarity values among all pairs of locations belonging to the same cluster (intra-cluster similarities); then, we did the same for the pairs belonging to different clusters (inter-cluster similarities). After that, we applied a kernel smoothing technique known as Kernel Density Estimation [22] to estimate the probability density functions for our similarity distributions, plotted in Figure 5 (the distribution of each cluster is represented by its color corresponding to Table 2).

We applied a t-test to determine if any given pair of distributions of intra- and inter-cluster similarity might originate from the same distribution, assessing that all distributions (and, therefore, the clusters) are significant at the 99% confidence level.

Table 2: Clusters of cities according to trend similarity.

Green	Yellow	Red	Purple
Long Beach	Memphis	St. Louis	Washington
Fresno	Salt Lake City	San Antonio	New York
Mesa	Harrisburg	Milwaukee	Detroit
Tucson	New Orleans	Tampa	Boston
Albuquerque	Baton Rouge	Pittsburgh	San Francisco
Virginia Beach	Portland	New Haven	Cleveland
San Jose	Tallahassee	Seattle	Minneapolis
Colorado Springs	San Diego	Cincinnati	Las Vegas
Jackson	Kansas City	Austin	Houston
Honolulu	Oklahoma City	Orlando	Charlotte
El Paso	Birmingham	Baltimore	Raleigh
Omaha	Louisville	Greensboro	Los Angeles
	Jacksonville	Nashville	Dallas-Ft. Worth
		Norfolk	Chicago
		Providence	Philadelphia
		Denver	Miami
		Richmond	Atlanta
		Phoenix	
		Sacramento	
		Columbus	
		Indianapolis	

We also compared the result of the hierarchical clustering with that of two network clustering algorithms (namely, Infomap [42] and the 'Louvain method' [4]) applied to the trend pathway backbone network (described in §2.2). We obtained consistent results in all cases: the only difference

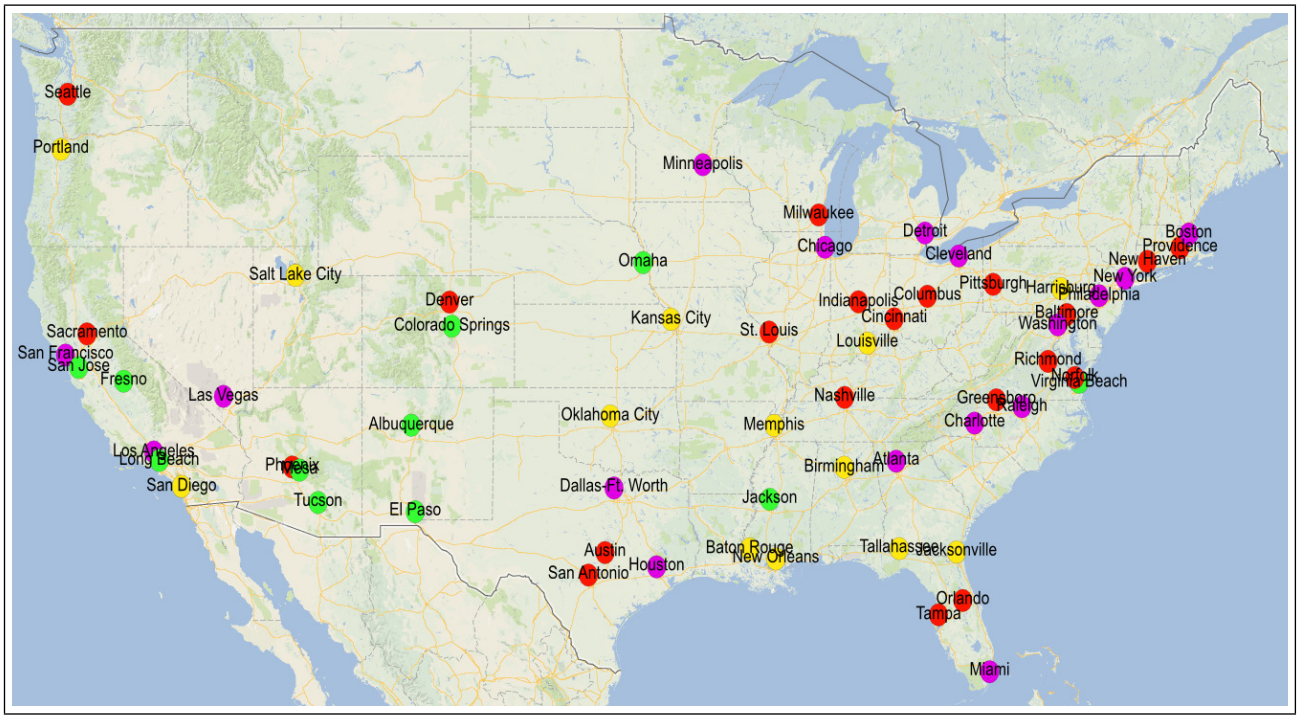

Figure 4: geographic representation of the 63 locations and respective clusters.

was that Seattle was placed in the purple cluster by both network clustering methods.

3.3 Trend pathway analysis

To establish where trends start and what pathways they follow to diffuse in the country, we analyze the multiscale trend pathway backbone network, built as described in §2.2 and represented in Figure 6 by using a divided edge bundling technique [47]. This visualization strategy has been successfully applied to other geographic networks such as the US airport traffic network (*cf.* [47]). In this node-link representation the edges are bundled taking into account directions and weights. The thicker the bundle, the higher the sum of the weights of connections wrapped in the bundle. In our case, this yields a network visualization that highlights the pathways followed by trends as they flow across

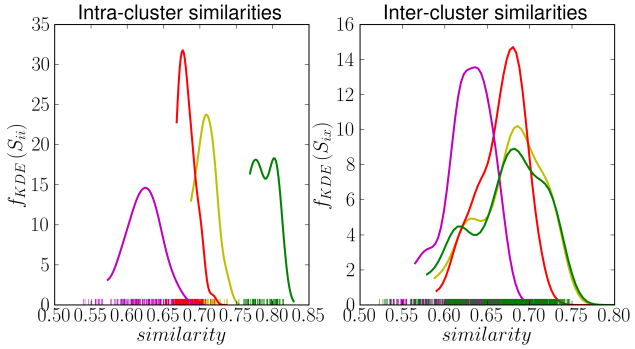

Figure 5: Kernel Density Estimation of intra- and inter-cluster similarity of the four clusters.

the country. In this figure the direction of edges represents the information flow: the tails of the bundles (in blue) show where trends start, the heads of the bundles (in red) point to where the trends arrive. From Figure 6 we can draw two observations: first, the presence of a massive backbone that carries the trend flow from the East coast to the West coast and vice-versa. Second, we observe a negligible North-South flow, except for that connecting Florida to the East coast. Moreover, the fact that the East-to-West flow is well balanced by the that in the opposite direction suggests that we are not simply observing an artifact of the time-zone effect: the West coast contributes to shaping the country trends to a similar extent that the East coast does.

In the backbone network the cities that often generate trends are those with higher fractions of outgoing edges (that is, those that spread their trends to most of the other cities); henceforth we will call them *sources*. Vice-versa, we will call *sinks* those cities with higher fraction of incoming edges. More precisely, since the network we deal with is weighted, we compute the *weighted source-sink ratio* $\omega(n)$ for each node n as

$$\omega(n) = \frac{s_{out}(n)}{s_{in}(n) + s_{out}(n)}, \tag{3}$$

where $s_{in}(n)$ (resp., $s_{out}(n)$) is the in-strength (resp., out-strength) of that node. We report in Table 3 the top 5 sources and the top 5 sinks of the backbone network. Four out of the five top sources (all but Cincinnati) also happen to be major metropolitan areas. On the other hand, all sinks belong to the Southwest and Midwest parts of the country. Los Angeles and New York (among our top sources) have also been reported in the top 5 hashtag producers worldwide in the recent work by Kamath *et al.* [23].

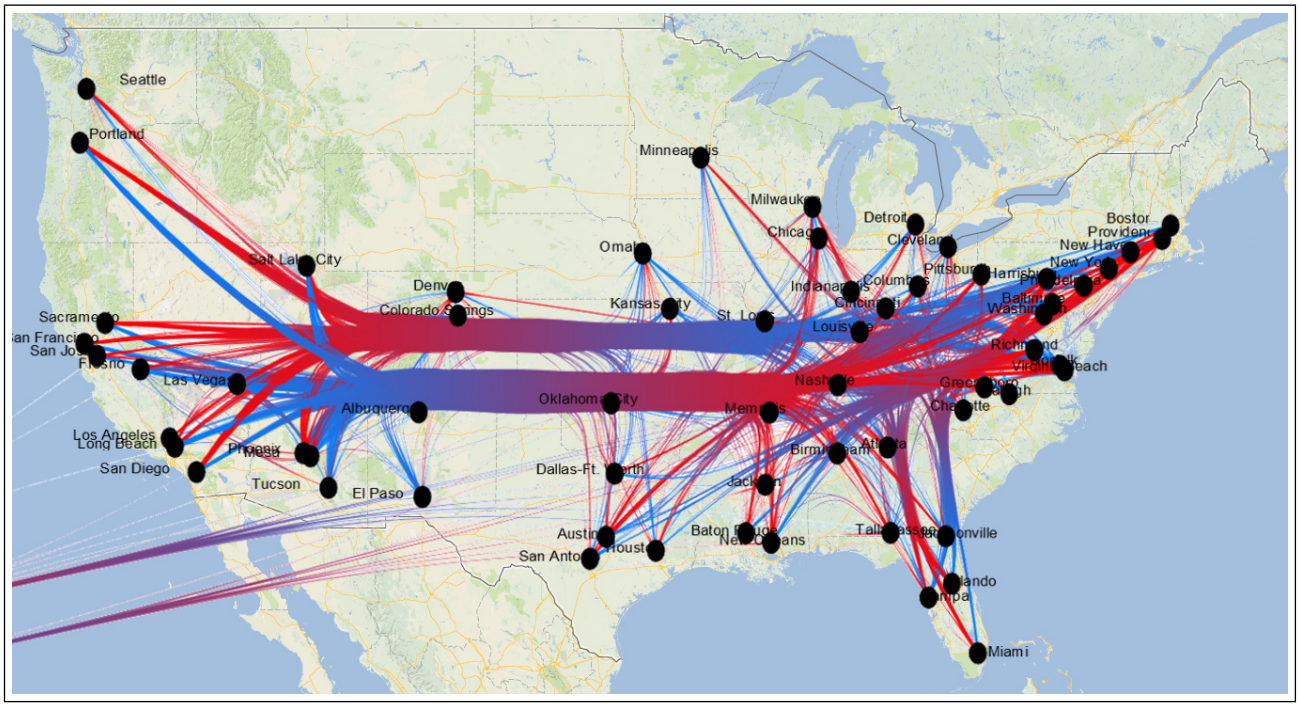

Figure 6: Trend pathways in Twitter. Trends spread in the direction from blue to red.

3.4 Trendsetters and trend-followers

The source-sink analysis presented above triggered our interest in the dynamics of trend popularity. In the following we study trendsetting and trend-following patterns, driven by the following question: *Are trending topics that become popular at the country level produced uniformly by all cities, or preferentially by some of them?*

To answer this question we selected from our dataset all those trends that at some point in time became trending at the country level. This left us with 1,724 hashtags and 2,768 phrases that achieved the highest popularity in the United States, appearing in the top 10 trending topics at the country level. We then selected the set of cities that exhibited each of these trends, and divided them in two categories: those cities in which the hashtag or phrase was trending *before* it became trending at the country level, and those cities that adopted it *after* it became trending at the country level. This allows us to determine what are the cities that contribute more to shaping the trends at the country level, and what are the cities that are more influenced by these global trends: in other words, we can identify trendsetters and trend-followers.

Figure 7 shows the result of this analysis for the hashtags. We can immediately identify two different classes of cities: the majority of them (*i.e.*, all those in the upper-left part of the main plot) appear to influence country-level trends roughly to the same extent to which they are influenced by the global trends; a second class of cities seem to have a much stronger trendsetting role toward the country.

To assess if these two classes can be significantly distinguished, we use the Expectation Maximization algorithm to learn an optimal Gaussian Mixture Model (GMM); to determine the appropriate number of components of the mixture we perform a 5-fold cross-validation using Bayesian and Akaike information criteria as quality measures, by varying the number of components from 1 to 10. The outcome of the cross-validation determines that the optimal number of components is two, according to both criteria, matching our expectations.

The result of the GMM is showed in the inset of Figure 7: each point is assigned to one of the two components yielding two different clusters composed respectively of 11 trendsetting cities (red dots) and 52 trend-following cities (blue stars). The list of trendsetters includes (in ascending order of impact) Raleigh, Detroit, Philadelphia, Houston, New York, Dallas-Ft. Worth, Boston, Denver, Atlanta, Los Angeles, and Seattle. All of them are major metropolitan areas.

To highlight the existence of these two different dynamics we applied a regression analysis approach by fitting two different linear regressions to the points belonging to the classes of trendsetters (coefficient of determination $R^2 = 0.9455$, p-value $p = 3.9 \cdot 10^{-7}$) and trend-followers ($R^2 = 0.7063$, $p < 10^{-10}$). This points out the proportionality that exists between incoming and outgoing trend flows.

We repeated this analysis by making the model even more realistic: for example, we introduced the effect of the time lag, discounting the reward given to those cities that adopt a

Table 3: Left: top 5 sources (*i.e.*, trendsetters). Right: top 5 sinks (*i.e.*, trend-followers).

Location	Rank	$\omega(n)$	Location	Rank	$\omega(n)$
Los Angeles	1^{st}	0.806	Oklahoma City	63^{rd}	0.101
Cincinnati	2^{nd}	0.736	Albuquerque	62^{nd}	0.109
Washington	3^{rd}	0.718	El Paso	61^{st}	0.235
Seattle	4^{th}	0.711	Omaha	60^{th}	0.305
New York	5^{th}	0.669	Kansas City	59^{th}	0.352

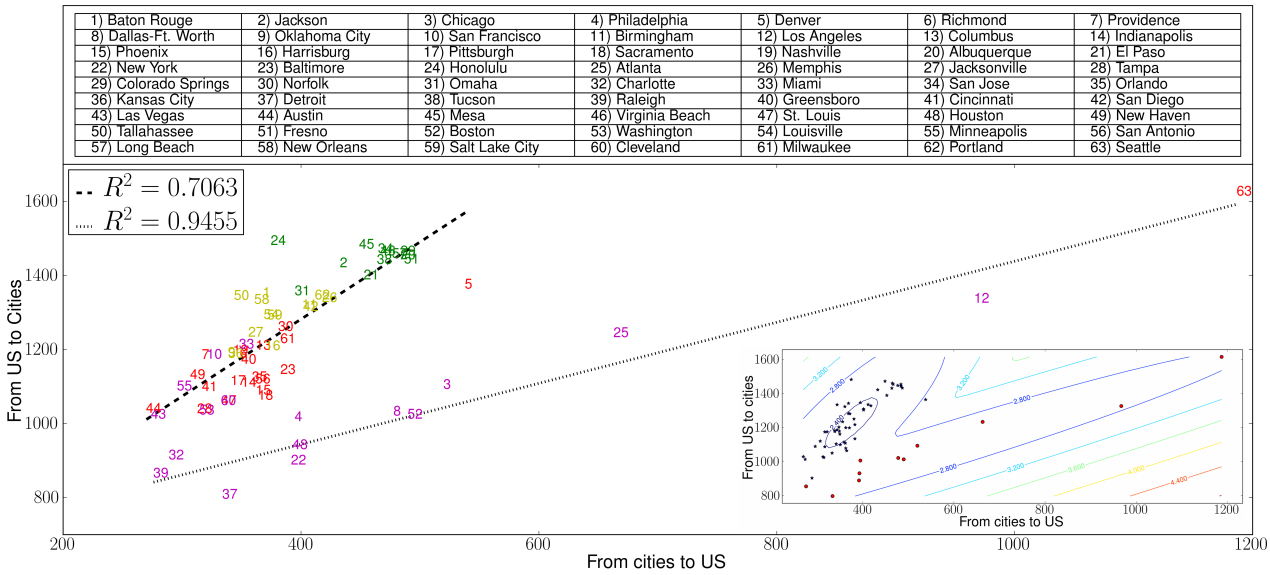

Figure 7: Trendsetting vs. trend-following cities. The x-axis shows the number of times a topic trending in a particular city later trends at the country level, while the y-axis shows the number of times of the reverse effect. The inset shows a Gaussian Mixture Model highlighting the two different trendsetting dynamics; the contours represent the standard deviations of each Gaussian distribution. In the main plot, two linear regressions are reported with the corresponding coefficient of determination R^2. City colors correspond to the cluster assignment in Table 2.

trend later with respect to the initiators; also, we rewarded only the initiators of each trend, rather than any city that exhibits a given trend before the trending point at the country level. Making the scenario more realistic did not affect the outcome: in all cases we obtained comparable results.

4. DISCUSSION

The fourth, purple cluster identified in §3.2 deserves further discussion. Differently from the others, this cluster is not geographically well defined (*cf.* Figure 4) — it contains metropolitan areas spread all over the country. Is the effect of city size sufficient to explain why these metropolitan areas are more influential than others, in the sense that they produce more national trends? It is not obvious that large populations would lead to more national trends: while a larger city produces more tweets and possibly more topic competing for popularity, the number of trends for each city at a given time is bounded to ten, irrespective of the city size. In cities with larger content production, hashtags (or phrases) must appear in more tweets to be listed as a trend, whereas a lower number of tweets is sufficient in cities with smaller content production. As a result, the effect of sheer volume is discounted by construction in the definition of Twitter trends.

Why, then, do the metropolitan areas in the purple cluster play such a trendsetting role? A possible interpretation is offered by noticing the presence in this cluster of some of the major airport hubs of the United States, such as Atlanta, Chicago, and Los Angeles. The list of top US airport hubs [52] is shown in Table 4, where we aggregated the traffic by metropolitan area. Surprisingly, 16 out of the 17 locations that constitute the cluster appear in the top 20 air traffic hubs — all of them but Cleveland. On the other hand, some cities in the cluster that do not belong in the top 30 metropolitan areas by population (Charlotte, Raleigh, Las Vegas), do appear among the major air traffic hubs.

Table 4: Top 20 cities ranked according to the total volume of flight traffic.

City	Cluster	Rank	Total traffic
New York (JFK, EWR, LGA)	purple	6^{th}, 14^{th}, 20^{th}	54,374,758*
Atlanta (ATL)	purple	1^{st}	45,798,809
Chicago (ORD, MDW)	purple	2^{nd}, 25^{th}	41,603,539*
Miami (MIA, FLL, PBI)	purple	12^{th}, 21^{st}, 54^{th}	33,228,913*
Dallas-Ft. Worth (DFW, DAL)	purple	4^{th}, 45^{th}	31,925,398*
Washington (BWI, IAD, DCA)	purple	22^{nd}, 23^{rd}, 26^{th}	31,431,854*
Los Angeles (LAX)	purple	3^{rd}	31,326,268
Denver (DEN)	red	5^{th}	25,799,832
Charlotte/Raleigh (CLT, RDU)	purple	8^{th}, 37^{th}	24,521,523*
Houston (IAH, HOU)	purple	11^{th}, 32^{nd}	24,082,666*
San Francisco (SFO)	purple	7^{th}	21,284,224
Las Vegas (LAS)	purple	9^{th}	19,941,173
Phoenix (PHX)	red	10^{th}	19,556,189
Orlando (MCO)	red	13^{th}	17,159,425
Seattle (SEA)	red	15^{th}	16,121,123
Minneapolis (MSP)	purple	16^{th}	15,943,751
Detroit (DTW)	purple	17^{th}	15,599,877
Philadelphia (PHL)	purple	18^{th}	14,587,631
Boston (BOS)	purple	19^{th}	14,293,675
Salt Lake City (SLC)	yellow	24^{th}	9,579,836

(*) Sum of the traffic volume of different airports in the same area.

The presence of major air traffic hubs among the special class of cities that act as trendsetters suggests an intriguing conjecture, drawing a parallel with the spread of diseases: *Does information travel faster by airplane than over the Internet?* In other words, do conversations and trends spread following social interaction dynamics, like *social butterflies*

that pass from person to person at the local level, or do they diffuse using traveling people as vectors, similarly to epidemics that take advantage of human mobility [13, 2]?

Further work is needed to explore this conjecture. One possibility would be to measure the correlation between trend overlap among pairs of cities and the corresponding air traffic.

5. RELATED WORK

Trends or aspects related to geography in socio-technical systems have been studied, directly or indirectly, in many recent studies. The present work is the first, to the best of our knowledge, that investigates the dynamics tightly binding trends and geography in online social media.

Geographic locations and physical distances have been found to be correlated to friendship behaviors in online social networks [28], to determine patterns in human mobility networks [7, 20], and to affect collaboration schemes in science networks [37].

Recent studies took advantage of platforms such as Yelp and Foursquare, which provide customized services to their users based on their physical location (*e.g.*, recommendations of events or places), to study geographic user activity patterns [35, 44, 45, 46].

Others have used platforms such as Twitter and Facebook, that enrich user profiles with geographic information and accompany user generated content with location-based data, to map users demographics [24, 31].

Onnela *et al.* [36] noted that, although the probability of observing a tie between two individuals in a social network (in that case, a mobile phone call network) decreases as a power law with physical distance, the geographic spread of social groups quickly increases with the size of the group; even groups of modest dimensions ($\approx$ 30 members) span across hundreds of kilometers, suggesting that, in technologically-mediated social systems, there exist distinctive social dynamics that govern the communication among individuals.

The findings presented in this paper nicely dovetail with Onnela's work, in that we observe the existence of a class of cities, geographically spread across the country, that acts as trendsetters for all other locations. On the other hand, we highlight that also a locality effect exists: geographically concentrated areas share similar contents and trends.

The local versus global ("glocal") nature of communication has been observed before in other types of online conversation [21]. In our analysis of the Occupy Wall Street movement on Twitter [14, 15], we noted that geographically localized discussions aim at mobilizing resources (*e.g.*, marshaling financial, material and human capital) while global discourse sets the goals of the movement and develops the narrative frames that reinforce collective purpose.

The influence of the locality effect has been also recently pointed out for innovation adoption on Twitter: Toole *et al.* [49] noted that homophily and physical closeness facilitate the adoption of new technological artifacts, suggesting that the effect of geographic location is critical to describe social dynamics in networked systems.

Geographic factors have also been recently found crucial in the adoption of languages and dialects [33], and in the expression of sentiment [32, 38, 39] in online social media. Mocanu *et al.* [33] showed how social media data can be used to characterize language geography at different levels of granularity, to highlight patterns such as linguistic homogeneity and linguistic mixture in multilingual regions.

Similarly, the study by Mitchell *et al.* [32] suggests that the adoption of online social media content can be instrumental to describe emotional, demographic and geographic characteristics of users of these socio-technical systems; in particular, they investigated Twitter users active in the US in terms of happiness and individual satisfaction.

Another recent research line related to our work is that of the detection of emerging trends, topics, memes, and events in online social networks and social media [1, 3, 10, 17, 19, 27, 30, 43]. Naaman *et al.* [34] characterized trends according to different dimensions, such as content, interaction, time-based and social features. These features were later used to classify trends, allowing for the identification of exogenous vs. endogenous trends and memes vs. retweet trends. In their analysis, the authors did not consider the geographic dimension, that is instead central in this work suggesting that it provides crucial information to characterize trends on online social media.

Finally, social media data can be used to make educated guesses on the outcome of real-word events, such as elections or competitions [18]. Ciulla *et al.* [12] combined trends and geographic information of Twitter data to demonstrate that online social media can be exploited to predict social events in the real-world. They collected trending hashtag and phrases related to contestants of the popular TV show *American Idol,* mapping the fan base of each candidate to different geographic regions inside and outside the US, to identify spatial patterns in attention allocation and preferences expressed on the online platform. These signals were then combined and used to predict voting behaviors of fans, achieving good accuracy.

6. CONCLUSIONS

In this work we investigated the spatial and geographic dynamics that govern trending topics in Twitter. We monitored trends from 63 different locations in the United States and, in addition, the trends at the country level, for a period of 50 days.

We sought to understand how trends are distributed in space and time and how they spread from place to place. We investigated shared trends among cities, finding that there exists a locality effect whose presence allows for the identification of three broad geographic areas where trends diffuse locally more than globally. We also identified a fourth cluster of metropolitan areas that counterbalances this locality effect. These cities, spread all over the country, act as sources of trends for other locations. They contribute much more than the others to shaping the global trends at the country level. We finally observed that these metropolitan areas coincide with the major air traffic hubs of the country, suggesting an intriguing conjecture based on a parallel between the spread of information and diseases: Do trends travel faster by airplane than over the Internet?

Our findings have broad potential applications, that include tailoring online content based on users geographic information, or designing better algorithms for geographic-aware trend prediction.

As for the future, our analysis opens new research questions that will need further attention. An example is the role of traffic hubs in trend diffusion. More in general, additional work is needed to understand how to identify locations that

can be influential for the spread of a given topic and how to effectively convey the information flow to determine the success of a given commercial campaign.

Acknowledgments

This work is supported by NSF (grant CCF-1101743), DARPA (grant W911NF-12-1-0037), and the McDonnell Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

7. REFERENCES

[1] C. Aggarwal and K. Subbian. Event detection in social streams. In *Proceedings of SIAM International Conference on Data Mining*, 2012.

[2] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489, 2009.

[3] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on Twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[5] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[6] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 450–453, 2011.

[7] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[8] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*, pages 241–250. ACM, 2012.

[9] C. Budak, D. Agrawal, and A. El Abbadi. Structural trend analysis for online social networks. *Proceedings of the VLDB Endowment*, 4(10):646–656, 2011.

[10] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the 10th International Workshop on Multimedia Data Mining*, page 4. ACM, 2010.

[11] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 2010.

[12] F. Ciulla, D. Mocanu, A. Baronchelli, B. Gonçalves, N. Perra, and A. Vespignani. Beating the news using social media: the case study of American Idol. *EPJ Data Science*, 1(1):1–11, 2012.

[13] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015–2020, 2006.

[14] M. D. Conover, C. Davis, E. Ferrara, K. McKelvey, F. Menczer, and A. Flammini. The geospatial characteristics of a social movement communication network. *PloS ONE*, 8(3):e55957, 2013.

[15] M. D. Conover, E. Ferrara, F. Menczer, and A. Flammini. The digital evolution of Occupy Wall Street. *PloS ONE*, 8(5):e64679, 2013.

[16] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. Political polarization on Twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[17] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski. #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17:124–147, 2013.

[18] J. DiGrazia, K. McKelvey, J. Bollen, and F. Rojas. More tweets, more votes: Social media as a quantitative indicator of political behavior. *Available at SSRN: http://dx.doi.org/10.2139/ssrn.2235423*, 2013. Presented at 108th annual meeting of the American Sociological Association.

[19] E. Ferrara, M. JafariAsbagh, O. Varol, V. Qazvinian, F. Menczer, and A. Flammini. Clustering memes in social media. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2013.

[20] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[21] K. Hampton and B. Wellman. Neighboring in netville: How the internet supports community and social capital in a wired suburb. *City & Community*, 2(4):277–311, 2003.

[22] T. Hastie, R. Tibshirani, and J. J. H. Friedman. *The elements of statistical learning*. Springer New York, 2001.

[23] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 667–677, 2013.

[24] J. Kulshrestha, F. Kooti, A. Nikravesh, and K. P. Gummadi. Geographic dissection of the Twitter network. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[25] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600. ACM, 2010.

[26] J. Lehmann, B. Gonçalves, J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in Twitter. In *Proceedings of the 21st International Conference on World Wide Web*, pages 251–260, 2012.

[27] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In

Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 497–506. ACM, 2009.

[28] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[29] A. Marcus, M. Bernstein, O. Badar, D. Karger, S. Madden, and R. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, pages 227–236. ACM, 2011.

[30] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the Twitter stream. In *Proceedings of the 2010 International Conference on Management of Data*, pages 1155–1158. ACM, 2010.

[31] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the demographics of Twitter users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[32] L. Mitchell, K. D. Harris, M. R. Frank, P. S. Dodds, and C. M. Danforth. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PloS ONE*, 8(5):e64417, 2013.

[33] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The Twitter of Babel: Mapping world languages through microblogging platforms. *PloS ONE*, 8(4):e61981, Jan. 2013.

[34] M. Naaman, H. Becker, and L. Gravano. Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918, 2011.

[35] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media)*, 2011.

[36] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis. Geographic constraints on social network groups. *PLoS ONE*, 6(4):e16939, 2011.

[37] R. K. Pan, K. Kaski, and S. Fortunato. World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports*, 2, 2012.

[38] D. Quercia. Don't worry, be happy: The geography of happiness on facebook. In *Proceedings of ACM Web Science 2013*, 2013.

[39] D. Quercia, L. Capra, and J. Crowcroft. The social world of Twitter: Topics, geography, and emotions. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[40] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[41] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*, pages 249–252. ACM, 2011.

[42] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–1123, 2008.

[43] H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams. In *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*, 2009.

[44] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft. Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades. In *Proceedings of the 20th International Conference on World Wide Web*, pages 457–466. ACM, 2011.

[45] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Distance matters: geo-social metrics for online social networks. *Proceedings of the 3rd Workshop on Online Social Networks*, 10, 2010.

[46] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media)*, pages 329–336, 2011.

[47] D. Selassie, B. Heller, and J. Heer. Divided edge bundling for directional network data. *IEEE Trans. Visualization & Comp. Graphics*, 17:2354–2363, 2011.

[48] M. Á. Serrano, M. Boguñá, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(16):6483–6488, 2009.

[49] J. L. Toole, M. Cha, and M. C. González. Modeling the adoption of innovations in the presence of geographic and media influences. *PloS ONE*, 7(1):e29528, 2012.

[50] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2, 2012.

[51] Wikipedia. Cities and metropolitan areas of the United States. `http://en.wikipedia.org/wiki/Cities_and_metropolitan_areas_of_the_United_States`, 2012.

[52] Wikipedia. List of the busiest airports in the United States. `http://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States`, 2012.

[53] S. Wu, J. Hofman, W. Mason, and D. Watts. Who says what to whom on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 705–714. ACM, 2011.

[54] L. Xie, A. Natsev, J. R. Kender, M. Hill, and J. R. Smith. Visual memes in social media: tracking real-world news in Youtube videos. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 53–62. ACM, 2011.