Spatio-Temporal Dynamics of Online Memes: A Study of Geo-Tagged Tweets

Krishna Y. Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng
Department of Computer Science and Engineering, Texas A&M University
College Station, TX, USA - 77843
{kykamath, caverlee, kyumin, zcheng}@cse.tamu.edu

ABSTRACT

We conduct a study of the spatio-temporal dynamics of Twitter hashtags through a sample of 2 billion geo-tagged tweets. In our analysis, we (i) examine the impact of location, time, and distance on the adoption of hashtags, which is important for understanding meme diffusion and information propagation; (ii) examine the spatial propagation of hashtags through their focus, entropy, and spread; and (iii) present two methods that leverage the spatio-temporal propagation of hashtags to characterize locations. Based on this study, we find that although hashtags are a global phenomenon, the physical distance between locations is a strong constraint on the adoption of hashtags, both in terms of the hashtags shared between locations and in the timing of when these hashtags are adopted. We find both spatial and temporal locality as most hashtags spread over small geographical areas but at high speeds. We also find that hashtags are mostly a local phenomenon with long-tailed life spans. These (and other) findings have important implications for a variety of systems and applications, including targeted advertising, location-based services, social media search, and content delivery networks.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*

Keywords

social media; spatial impact; spatio-temporal analysis

1. INTRODUCTION

The rise of social media services enables a global-scale infrastructure for the sharing of videos, blogs, images, tweets, and other user-generated content. As users consume and share this content, some content may gain traction and become popular resulting in viral videos and popular memes that captivate the attention of huge numbers of users. These phenomena have attracted a considerable amount of recent research to study the dynamics of the adoption of social media, e.g., [4, 13, 15, 17, 20].

Augmenting this rich body of research is the widespread adoption of GPS-enabled tagging of social media content via smartphones and social media services, which provides new access to the fine-grained spatio-temporal logs of user activities. For example, the Foursquare location sharing service has enabled 2 billion "check-ins" [12], whereby users can link their presence, notes, and photographs to a particular venue. The mobile image sharing service Instagram allows users to selectively attach their latitude-longitude coordinates to each photograph; similar geo-tagged image sharing services are provided by Flickr and a host of other services. And the popular Twitter service sees ~300 million Tweets per day, of which ~3 million are tagged with latitude-longitude coordinates.

Access to these geo-spatial footprints opens new opportunities to investigate the spatio-temporal dynamics of online memes, which has important implications for a variety of systems and applications, including targeted advertising, location-based services, social media search, and content delivery networks. Hence, in this paper, we initiate a study of the *spatio-temporal properties* of social media spread through an examination of the fine-grained sharing of one type of global-scale social media – a sample of 2 billion geo-tagged Tweets with precise latitude-longitude coordinates collected over the course of 18 months. Specifically we consider the *propagation* of hashtags across Twitter, where a hashtag is a simple user-generated annotation prefixed with a #. Hashtags serve many purposes on Twitter, from associating Tweets with particular events (e.g., #ripstevejobs and #fukushima) to sharing memes and conversations (e.g., #bestsportsrivalry and #ifyouknowmeyouknow). Our goal is to explore questions such as:

- What role does distance play in the adoption of hashtags? Does distance between two locations influence both what users in different locations adopt and when they do so?
- While social media is widely reported in terms of viral and global phenomenon, to what degree are hashtags truly a global phenomenon?
- What are the geo-spatial properties of hashtag spread? How do local and global hashtags differ?
- How fast do hashtags peak after being introduced? And what are the geo-spatial factors impacting the timing of this peak?
- How can the spatio-temporal characteristics of hashtags describe locations? Are some locations more "impactful" in terms of the hashtags that originate there?

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

While limited to one type of social media spread and with an inherent sample bias towards using who are willing to share their precise location, the investigation of these questions can provide new insights toward understanding the spatio-temporal dynamics of the sharing of user-generated content. Our investigation is structured in three steps. First, we study the global footprint of hashtags and explore the spatial constraints on hashtag adoption. In particular, we analyze the worldwide distribution of hashtags and the impact distance has on where and when hashtags will be adopted. Second, we study three spatial properties of hashtag propagation – focus, entropy, and spread – and examine the spatial propagation of hashtags using these properties. Specifically, we study the nature (local or global) of hashtag propagation and the correlation between the spatial properties and the number of occurrences of the hashtags. Finally, we present two methods for characterizing locations based on hashtag spatial analytics. The first method uses spatial properties – entropy and focus – to determine the nature of a location from the point of hashtag propagation, while the second method uses hashtag adoption times to characterize a location's impact to enable hashtag propagation.

Some of our key findings are:

- Hashtags are a global phenomenon, with locations all across the world. But the physical distance between locations is a strong constraint on the adoption of hashtags, both in terms of the hashtags shared between locations and in the timing of when these hashtags are adopted.
- Hashtags are essentially a local phenomenon with long-tailed life spans, but follow a "spray-and-diffuse" pattern, similar to Youtube videos [5], where initially a small number of locations "champion" a hashtag, make it popular, and the spread it to other locations. After this initial spread, hashtag popularity drops and only locations that championed it originally continue to post it.
- The rate at which a hashtag becomes popular is dependent on the hashtag's origin. That is, hashtags that originate as responses to external stimuli (like real-world events) spread faster than hashtags that originate purely within the Twitter network itself (e.g., corresponding to a Twitter meme like #ifyouknowmeyouknow).
- Through hashtag spatial analytics, the relative impact of locations can be measured; for example while both London and Sao Paulo are home to the most total hashtags, hashtags originating in London have a global footprint, while Sao Paulo's are mostly constrained to Brazil due to inherent language and culture constraints.

These results can positively impact both research into the spread of online memes as well as systems operators, e.g., informing the design of distributed content delivery networks and search infrastructure for real-time Twitter-like content. For example, caching decisions to improve fast delivery of social media content to users and to support applications like real-time search can build upon the results presented here. Insights into the role distance plays and the impact locations have on hashtag spread could inform new algorithms for geo-targeted advertising. This work can also complement efforts to model network structures that support (or impede) the "viralness" of social media, measure the contagion factors that impact how users influence their neighbors, develop models of future social media adoption, and so forth.

2. RELATED WORK

Our work presented here builds on two lines of research: studies of Twitter and of Twitter hashtags; and geo-spatial analysis of social media.

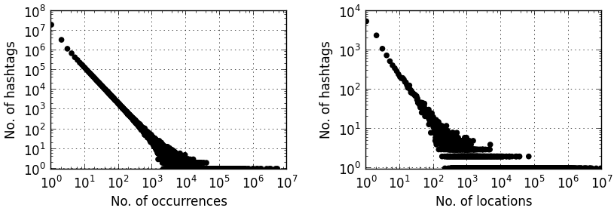
Twitter Hashtag Analysis: There have been several papers studying the general properties of Twitter as a social network and in analyzing information diffusion over this network [13, 15, 23, 16]. Continuing in this direction most papers related to hashtags have focused their attention on understanding the propagation of hashtags on the network. For example, in [20] the authors studied factors for hashtag diffusion and found that repeated exposure to a hashtag increased the chance of it being reposted again, especially if the hashtag is contentious. An approach grounded in linguistic principles has studied the properties of hashtag creation, use, and dissemination in [8]. In related research, approaches based on linear regression have been used to predict the popularity of hashtags in a given time frame [22]. Because of the variety of ways in which hashtags's are used to convey information about a tweet, there has been recent research in hashtag-based sentiment detection [10], topic tracking on twitter streams [18], and so forth.

Geo-spatial Analysis of Social Media: The emergence of location-based social networks like Foursquare, Gowalla, and Google Latitude has motivated large-scale geo-spatial analysis [14, 21, 19, 7]. Some of the earliest research related to geo-spatial analysis of web content were based on mining geography specific content for search engines [11]. More recently in [2] the authors analyzed search queries to understand the spatial distribution of queries and understand their geographical centers. On Twitter, geo-spatial analysis has focused on inferring geographic information from tweets like predicting user locations from tweets [6] and spatial modeling to geolocate objects [9]. Similar analysis to infer a user's location on Facebook based on their social network has been studied in [3]. Researchers have also analyzed Youtube videos for geo-spatial properties and observed the highly-local nature of video views [5].

3. DATA AND SETUP

We collected a sample of 2 billion geo-tagged tweets containing 342 million hashtags (27 million unique hashtags) from Twitter using the Twitter Streaming API from February 1, 2011 to October 31, 2012. Each tweet in this sample is tagged with a latitude and longitude indicating the location of the user at the time of the posting, resulting in a tuple of the form $<$ `hashtag`, `time`, `latitude`, `longitude` $>$. The expected long tail distribution for hashtag occurrences is shown in Figure 1(a).

To support location-based analysis, we divide the globe into square grids of equal area using Universal Transverse Mercator (UTM), a geographic coordinate system which uses a 2-dimensional Cartesian coordinate system to map locations on the surface of the globe [1]. The issue with using an angular co-ordinate system like latitudes and longitudes



(a) Hashtag distribution (b) Location distribution

Figure 1: Hashtag dataset properties

is that distance covered by a degree of longitude differs as we move towards the pole. In addition, the distance covered by moving a degree in latitude and longitude is same only at the equator. Hence, it is hard to break globe into grids using this system. UTM on the other hand gives us a system of grids that closely matches distances in metric system making our analysis easier. While varying the choice of grid size can allow analysis at multiple levels (e.g., from state-sized cells to neighborhood-sized ones), we adopt a middle ground by dividing the globe into squares of 10km by 10km. Some grid cells will naturally be densely populated, others will be sparse. Let this set of distinct locations, each corresponding to a square, be represented by the set L. With these locations, we observe in Figure 1(b) that the number of hashtags present in a location follows a long tail distribution (e.g., 10,000 unique hashtags are observed in 10 locations; 100 unique hashtags are observed in 100 locations), following the expected population density of equal-sized grid cells.

For the rest of the paper, we focus on hashtags with at least 5 occurrences in a location and with at least 50 total occurrences across all locations. Since some hashtags may have begun their Twitter life before the first day of our sample while others may have continued on after the last day, we consider both February 2011 and October 2012 as buffer months. Hence, we capture the full lifecycle of hashtags starting on or after March 1, 2011 and ending by September 30, 2012 which focuses our study to hashtags which have both their birth and death within the time of study.

The rest of this paper considers a set of hashtags H (consisting of close to 20 million hashtags from 99,015 unique hashtags) and a set of locations L (consisting of 4,946 locations). For every hashtag ($h \in H$) and location ($l \in L$) pair, we denote the set of all occurrences of h in l as O_l^h. We say that H_l is the set of unique hashtags observed in l.

Our study continues in three major parts:

- First, we study the global footprint of hashtags and explore the spatial constraints on hashtag adoption. (§ 4)
- Second, we study three spatial properties of hashtag propagation – focus, entropy, and spread – and examine the spatial propagation of hashtags using these properties. (§ 5)
- Finally, we illustrate two potential methods for characterizing locations based on hashtag spatial analytics. (§ 6)

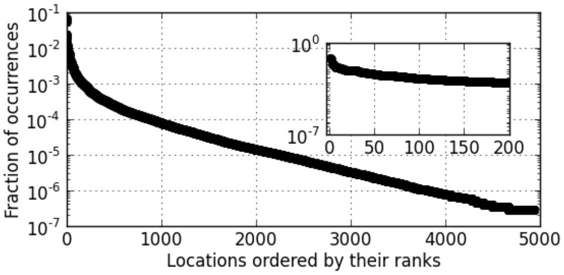


Figure 2: Fraction of hashtag occurrences in locations ordered by their rank. The inset plot shows the fraction for top-200 locations.

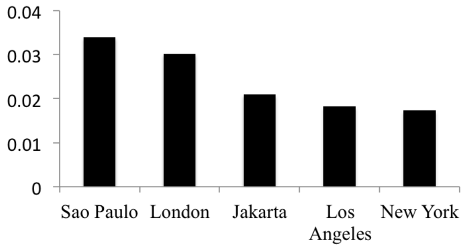


Figure 3: Top-5 locations with most hashtags

4. LOCATION PROPERTIES OF HASHTAGS

In this section, we begin our analysis by examining the locations represented in the dataset and exploring the relationship between locations. In particular, we are interested in understanding: (i) what is the worldwide distribution of hashtags? (ii) does distance between two locations influence which hashtags they adopt? and (iii) does distance between two locations influence when they will adopt these hashtags?

4.1 Location Distribution

We first examine the distribution of hashtags across the 4,946 unique locations represented in the dataset, as shown in Figure 2. The distribution of hashtags occurring in locations ordered by their rank (in terms of number of occurrences) decreases exponentially with increasing rank, meaning that the distribution of hashtags in various locations is very uneven. But, focusing on just the top-200 locations (as shown in the inset plot in Figure 2), we see that though the decrease in occurrence is exponential, it is small compared to the drop that we see for all locations in the larger figure, indicating the presence of locations which generate high but relatively the same number of hashtags.

The top-5 locations by their rank are shown in Figure 3. While Sao Paulo claims close to 3.4% of all hashtags and no US city occurs in the top-3 positions, when aggregating locations by country we observe that the US has close to a 40% share followed by Brazil with 6% and the UK with 5%.

Although the US dominates, if we extend to the top-200 most prevalent locations, we see in Figure 4 the global footprint of hashtags covering most of the major densely populated cities in the world (sans China).

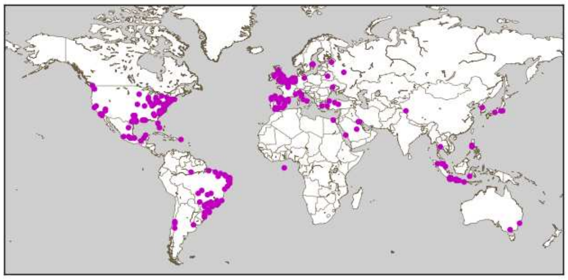


Figure 4: Top-200 locations with the most hashtags.

4.2 Relationship between Locations

Given the global nature of hashtags, we next examine the relationship between locations in terms of hashtag adoption. We consider two approaches that consider the distance between location pairs – one based on the fraction of hashtags shared between locations; the other based on the adoption time lag between locations. In both cases we measure the distance between locations using the Haversine distance function, which accounts for the effects of the Earth's spherical shape in finding distances between points.[1] In essence, the Haversine maps from latitude-longitude pairs to distance: $\mathcal{D} : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$.

Hashtag Sharing vs Distance: We first seek to understand the relationship of the distance between locations on the commonality of hashtags adopted in locations. To what degree does distance impact whether a hashtag is shared between two locations? Given two locations, we measure their hashtag "similarity" using the Jaccard coefficient between the sets of hashtags observed at each location:

$$\text{Hashtag Similarity}(l_i, l_j) = \frac{H_{l_i} \cap H_{l_j}}{H_{l_i} \cup H_{l_j}}$$

where recall H_l is the set of unique hashtags observed in l. Locations that have all hashtags in common have a similarity score of 1.0, while those that share no hashtags have a score of 0.0. The relationship between hashtag similarity and distance is plotted in Figure 5. We see a strong correlation, suggesting that the closer two locations are, the more likely they are to adopt the same hashtags. As distance increases, the hashtag sharing similarity drops accordingly. Much of this distance-based correlation can be explained by issues of language, culture and other common interests shared between these locations. For example, we see strong similarities in hashtags between English-speaking parts of Western Europe and the United States; and between Portuguese-speaking parts of Brazil and Portugal.

Hashtag Adoption Lag vs Distance: While locations that are near are more likely to share hashtags, are they also more likely to adopt hashtags at the same time? We next measure the impact of distance on hashtag adoption lag between two locations. Locations that adopt a common hashtag at the same time can be considered as more temporally similar than are two locations that are farther apart in time (with a greater lag). Letting t_l^h be the first time when hashtag h was observed in location l, we can define the hashtag adoption lag of two locations as:

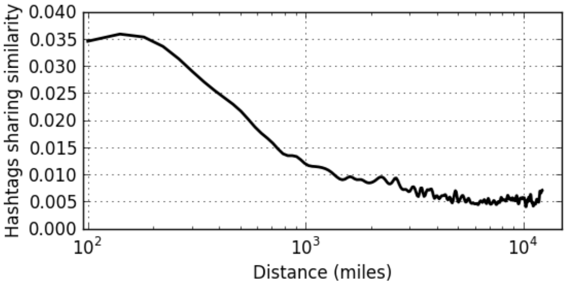


Figure 5: Hashtag sharing similarity vs Distance.

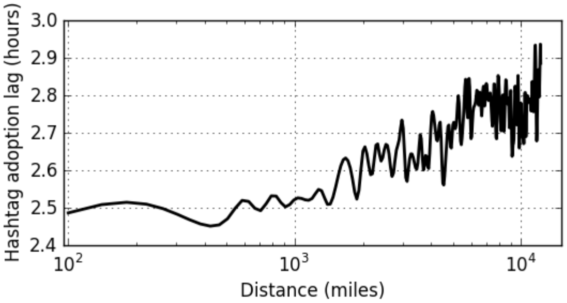


Figure 6: Hashtag adoption lag vs Distance

$$\text{Adoption Lag}(l_i, l_j) = \frac{1}{|H_{l_i} \cap H_{l_j}|} \sum_{h \in H_{l_i} \cap H_{l_j}} |t_{l_i}^h - t_{l_j}^h|$$

where the adoption lag measures the mean temporal lag between two locations for hashtags that occur in both the locations. A lower value indicates that common hashtags reach both the locations around the same time. We see in Figure 6 a relatively flat relationship up to ~500 miles, then a generally positive correlation, suggesting that locations that are close in spatial distance tend also to be close in time (e.g., they adopt hashtags at approximately the same time). Locations that are more spatially distant tend to adopt hashtags at greater lags with respect to each other.

4.3 Summary

Our observations in this section indicate that hashtags are fundamentally a global phenomenon, with locations all across the world participating in the sharing of this type of social media. However, we have also confirmed that the physical distance between locations is a strong constraint on the adoption of hashtags, both in terms of the hashtags shared between locations and in the timing of when these hashtags are adopted.

5. HASHTAG PROPAGATION

Based on the observations in the previous section, we now focus on the characteristics of hashtag propagations across the globe. We examine the spatio-temporal properties of individual hashtags to explore questions like: To what degree are hashtags a local phenomenon? Does the number of occurrences of hashtag impact its global spread? Can we

[1] For a fuller treatment, we refer the interested reader to http://en.wikipedia.org/wiki/Haversine_formula

characterize the spatial properties of local and global hashtags?

5.1 Spatial Properties of Hashtag Propagation

Previous studies of the geographic scope of social media and web resources have typically adopted two types of measures: one considering the intensity of focus and one considering the uniformity of this interest. Similarly, we adopt two measures (similar to ones for studying YouTube videos in [5]): *hashtag focus* and *hashtag entropy*, plus a third measure called the *hashtag spread*.

For every hashtag ($h \in H$) and location ($l \in L$) pair, if we let O_l^h be the set of all occurrences of h in l, then the probability of observing hashtag h in location l is defined as:

$$P_l^h = \frac{O_l^h}{\sum_{l \in L}\{O_l^h\}}$$

Then the *hashtag focus* for hashtag h is:

$$\mathcal{F}^h = \max_{l \in L} P_l^h$$

which is simply the maximum probability of observing the hashtag at a single location. The location at which the probability is maximum is called the *hashtag focus location*. As a hashtag propagates, intuitively its focus will reduce as the hashtag is observed at multiple locations. The more local a hashtag is, presumably the higher its focus will be as well. Note that we additionally denote the focus measured over an interval t (rather than over the entire dataset) as $\mathcal{F}^h(t)$.

The *hashtag entropy* is defined as:

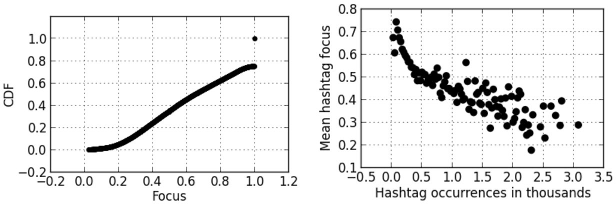
$$\mathcal{E}^h = -\sum_{l \in L} P_l^h \log_2 P_l^h$$

which measures the randomness in spatial distribution of a hashtag and determines the minimum number of bits required to represent the spread. A hashtag that occurs in only a single location will have an entropy of 0.0. As a hashtag spreads to more locations, its entropy will increase, reflecting the greater randomness in the distribution. Like focus, we can additionally denote the entropy measured over an interval t (rather than over the entire dataset) as $\mathcal{E}^h(t)$.

While focus and entropy provide insights into a hashtag's locality, they lack explicit consideration for the distance a hashtag has traveled. For example, consider two hashtags – one distributed equally between Austin and Dallas, and another one equally distributed between Los Angeles and New York. The focus of both hashtags is 0.5 and their entropy is 1. Hence, to measure the greater "dispersion" of the LA-NY hashtag, we define the *hashtag spread* of hashtag h as:

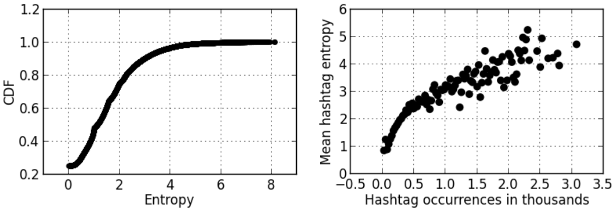
$$\mathcal{S}^h = \frac{1}{|O^h|} \sum_{o \in O^h} \mathcal{D}(o, G(O^h))$$

which measures the mean distance for all occurrences of a hashtag from its geographic midpoint. Here, G is the geographic midpoint[2] for a set of occurrences, which is similar to calculating the midpoint on a plane for a set of 2-dimensional points, but as in the case of Haversine distance, the geographic midpoint is calculated by considering the effects of Earth's spherical shape. A local hashtag with many occurrences close to its midpoint will yield a small spread, while a global hashtag with occurrences relatively far from its center will yield a larger spread.



(a) CDF (b) Mean hashtag focus

Figure 7: Focus: Around 50% of hashtags accumulate at least 50% of their postings from a single location.



(a) CDF (b) Mean hashtag entropies

Figure 8: Entropy: Almost 20% of hashtags are confined to a single location, but hashtags begin to spread as they become popular.

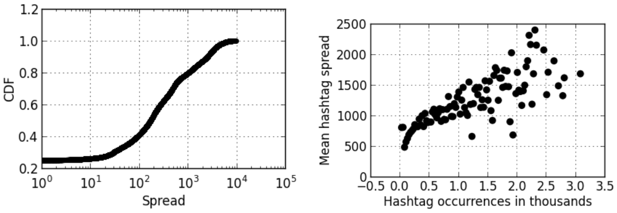
5.2 Local versus Global: Measuring Focus, Entropy, and Spread

Using these three spatial properties, we now analyze the properties of hashtag propagations.

Measuring Hashtag Focus: We begin by considering the focus values of hashtags. The cumulative distribution for focus values of hashtags is shown in Figure 7(a). We observe that the distribution is nearly linear, meaning that the focus values for hashtags are uniformly distributed. We also notice that most hashtags are concentrated in one location. Specifically, around 50% of hashtags derive at least 50% of their postings from a single location. In addition, as indicated by the single dot at CDF = 1.0, about a quarter of all hashtags are observed in a single location only. Continuing this look at hashtag focus, we next plot the relationship between the number of occurrences of a hashtag and its focus in Figure 7(b). As can be expected, we observe that hashtags with a few occurrences have a high focus (meaning that these low-intensity hashtags tend to occur primarily in a single location), whereas an increasing number of occurrences corresponds to a decrease in the focus of the hashtag. Together, these results suggest that many hashtags correspond to either local events (e.g., #momentoschampions, #nyadaauditions) or geographically compact networks of friends. But as hashtags become more popular they tend to spread to more locations. That is, it is unlikely for a popular hashtag to be constrained to a handful of locations; there is spillover from one location to the next.

Measuring Hashtag Entropy: To further explore this spatial distribution, we next consider the entropy of hashtag propagations. Recall that an entropy of zero for a hashtag indicates that it was posted from one (2^0) location only, while, for example, an entropy value of two indicates a hash-

[2] http://www.geomidpoint.com/



(a) CDF (b) Mean hashtag spread

Figure 9: Spread: 50% of hashtags have a spread less than 400 miles; 25% of hashtags have a spread greater than 1000 miles.

tag propagated almost equally to four (2^2) locations. The cumulative distribution of entropy in Figure 8(a) shows that about 25% of hashtags are concentrated in a single location and that the majority of hashtags propagate to at most two locations. On the flipside, however, we do see that hashtags with many occurrences tend to spread to many locations, as seen by the increasing entropy versus the number of hashtag occurrences in Figure 8(b) (and the decreasing focus values, as we observed in Figure 7(b)). As a hashtag becomes popular it tends to spread to newer locations and this in turn makes it more popular. These results show that the majority of hashtags have a narrow base of geographic support, but that one of the keys to popularity is a broad geographic footprint. This is intuitively sensible, but important to confirm in practice.

Measuring Hashtag Spread: While focus and entropy provide insights into a hashtag's locality, neither directly measures the geographic area over which a hashtag propagates. Using hashtag spread, we see in Figure 9(a) that about a quarter of hashtags have a spread of zero since they were observed in only location. In addition, we observe that most hashtags have a small spread, with almost 50% of hashtags having a spread less than 400 miles. However, we do observe that around 25% of hashtags have a spread greater than 1000 miles. We next plot the correlation between number of occurrences of a hashtag and its spread in Figure 9(b). Consistent with the findings over focus and entropy we observe that an increasing number of occurrences is coupled with a larger spatial footprint.

Direct Comparison of Spatial Properties: We now turn to directly comparing the focus, entropy, and spread values for our hashtags. We begin by plotting the mean hashtag focus on the x-axis versus the mean hashtag entropy on the y-axis, as shown in Figure 10. Local hashtags – with a high focus and a low entropy – are located in the bottom-right of the figure; global hashtags – with a low focus and a high entropy – are located in the top-left of the figure.

The correlation between spread and our two other spatial properties – focus and entropy – is shown in Figure 11. As expected, an increasing spread results in a decreasing focus because as a hashtag spreads it occurs in more locations which in turn reduces the overall focus. For similar reasons we observe an increase in entropy with increasing spread.

We also observe that in Figure 11(a), there is a steep drop in focus for the first 700 miles, followed by a region of almost uniform focus until about 1600 miles and finally a region of decreasing focus until 4000 miles. The initial steep drop of focus indicates that the locations that are adopting hashtags are spatially close to each other. On a map, the spatial distribution of these hashtags would look like a tight cluster of dots in a small region. The next region where the focus remains almost the same while the spread increases corresponds to hashtags that are spatially well distributed but the majority of hashtags are being produced by a single location. On a map the spatial distribution for these hashtags would have dots spread over a wide region as in Figure 12(a), but with only a few of those dots generating the majority of hashtags. Finally, the third region corresponds to globally distributed hashtags like the one shown in Figure 12(b). We see similar behavior when we plot entropy against spread as shown in Figure 11(b): a steep increase in entropy for the first 700 miles, then a region until about 1600 miles with uniform entropy and finally a region of increasing entropy until 4000 miles.

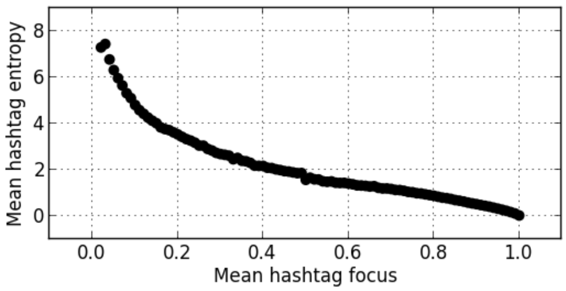
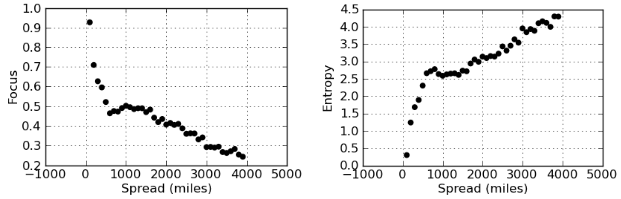


Figure 10: Entropy versus Focus.

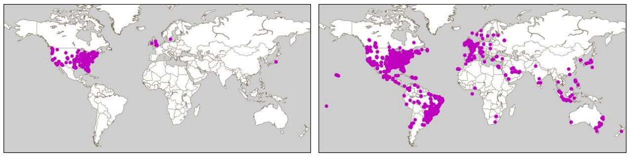


(a) Focus vs Spread. (b) Entropy vs Spread.

Figure 11: Correlation between spatial properties and spread.

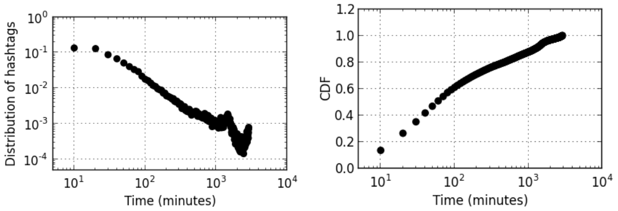
In summary, most hashtags are essentially a local phenomenon, as indicated by the on-average high focus, low entropy, and small spread. But as a hashtag becomes more popular, we see a decrease in focus and an increase in entropy and spread, all hallmarks of global impact. Based on the analysis in this section, we identify three broad categories of hashtags:

- **Local Interest [60% of all hashtags]**: These hashtags have a spread range from 0 to 700 miles. They have a high focus with median of 0.79 and low entropy of 1 bit. Example local interest hashtags include *#volunteer4betterindia*, *#ramadanmovies*, and *#onceuponatimeinnigeria*.
- **Regional and Event-Driven [15% of all hashtags]**: These hashtags have a spread range from 700 to 1000 miles. They have a median focus of 0.44 and



(a) #cnndebate (b) #ripstevejobs

Figure 12: Example of hashtag spread.



(a) Distribution of hashtag peaks. (b) CDF for hashtag peaks.

Figure 13: Hashtag peak analysis.

entropy of 3 bits. Example regional and event-driven hashtags include#cnbcdebate, #iowadebate, etc.

- **Worldwide Phenomena [25% of all hashtags]**: These hashtags have a spread range from 1000 to 4000 miles. These are mostly global hashtags which have low focus with median of 0.28 and entropy of 4 bits. Example worldwide phenomena hashtags include *#britneyvmas*, *#yearof4*, *#timessquareball*.

5.3 Slow versus Fast: Peak Analysis

We next augment our analysis by considering, in addition to the spatial propagation of hashtags, the temporal characteristics of these hashtags. We begin this temporal analysis by studying *when* hashtags reach the peak of their propagation in terms of occurrences. For this study we focus on hashtags that reach their peak within the first two days after their first appearance. We see in Figure 13(a) the distribution of *peak times* across all hashtags. We find that around 20% of hashtags reach their peak within 20 minutes of their first appearance. The distribution of peaks falls exponentially after that. We also observe that about 60% of all hashtags reach their peak within the first 2 hours as shown in Figure 13(b). In addition we observe that on average hashtags accumulate more than 50% of their total occurrences in the first 2 hours of their propagation as shown in Figure 14.

But what are the differences between fast-peaking hashtags and slow-peaking ones? Do hashtags behave differently in terms of their spatial properties? To answer these questions, we consider two sets of hashtags – those that reach their peak within the first 30 minutes of their initial appearance and a second set consisting of slower hashtags that reach their peak between 4 and 10 hours of their initial appearance. To analyze the relationship between locality and peak times we plotted these sets of hashtags in Figure 15, with focus on the x-axis and entropy on the y-axis.

We observe that in the set of faster hashtags – which reach a peak within 30 minutes of their propagation – the local hashtags are much faster than the global ones (see Figure 15(a)). This observation is reversed in the set of slower hashtags, shown in Figure 15(b), where the global hashtags are relatively faster than the local hashtags. On closer inspection, we attribute this reversal to the motive or purpose of the hashtags. First, we observe that hashtags that peak slowly are mostly of anticipated events, like the hashtag "#mtvema" corresponding to the MTV music awards, while the hashtags that peak more quickly are those that are organically generated within Twitter and related to fun like "#childhoodmemories". Second, slower hashtags are not as dependent on social sharing within Twitter as compared to faster hashtags; for example, users may be aware about the MTV awards from multiple sources (TV, news, friends), while the hashtag "#childhoodmemories" is seen only by those on Twitter. This dependency on the network to spread makes local fast hashtags peak sooner than the global fast hashtags. The global slow hashtags peak sooner than the local slow hashtags since more people are aware about them and they are not dependent on the network.

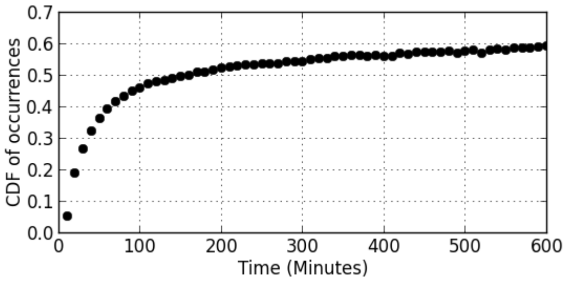
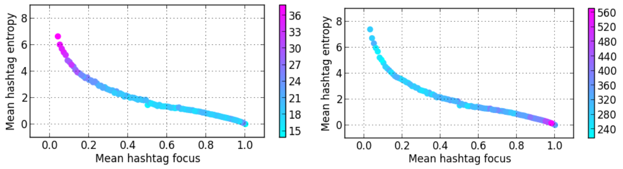


Figure 14: CDF of occurrences with time.



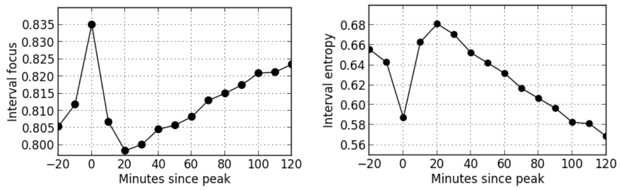
(a) Hashtags that peak during the first 30 minutes. (b) Hashtags that peak between 4 and 10 hours

Figure 15: (Color) Comparing the spatial properties of hashtags that reach their peak quickly (a) and those that reach their peak more slowly (b). Local hashtags – with a high focus and a low entropy – are located in the bottom-right of each figure; global hashtags – with a low focus and a high entropy – are located in the top-left of each figure. Low peak values are in light blue; high peak values in magenta.

Based on this peak analysis, we group hashtags into three categories:

- **Fast [25% of all hashtags]**: These hashtags reach their peak within 30 minutes of their first appearance. We find that 65% of these hashtags are local, 15% of these hashtags are national or event driven and 20% are global.

- **Medium [20% of all hashtags]**: These hashtags reach their peak between 30 minutes and 10 hours after their first appearance. We find that 55% of these



(a) Interval focus with time.

(b) Interval entropy with time.

Figure 16: Hashtags peak with their most "global" footprint 10-20 minutes after their peak

hashtags are local, 17% of these hashtags are national or event driven and 28% are global.

- **Slow [55% of all hashtags]**: These hashtags reach their peak more than 10 hours after their first appearance. We find that 60% of these hashtags are local, 16% of these hashtags are national or event driven and 24% are global.

For all three peak-based categories we observe that the distribution of spatial categories is quite similar.

5.4 Patterns of Hashtag Propagation

We next zoom in on the spatial properties of hashtag propagation during the minutes pre- and post- peak. When hashtags peak, do they peak suddenly in different locations simultaneously or do they slowly accumulate a larger spatial footprint? What are the dynamics of their spatial properties as they become popular?

For this study, we divide each hashtag's lifecycle into equal length time intervals of 10 minutes. For each time interval, we compute the hashtag focus ($\mathcal{F}^h(t)$) and the hashtag entropy ($\mathcal{E}^h(t)$) over just that interval. We plot these interval-specific focus and entropy measures in Figure 16. First, compared to the aggregate characteristics across all hashtags – in which we find the median focus for all hashtags over their entire lifetime to be 0.57; for entropy, we find a median of 2 bits – here we see that the interval-based focus is even higher (greater than 0.80 in all cases) and the interval-based entropy is even lower (less than 1 bit in all cases). These higher focus / lower entropy results indicate that hashtags are *even more local* during each step of their propagation. To illustrate, in the aggregate we may find a hashtag that propagates only in locations in Texas. Compared to a global hashtag, it is certainly more local and its focus and entropy will reflect this. However, during its propagation, the Texas-based hashtag is even more local at each step; that is, it does not propagate over the entire state simultaneously but in stages, city by city. It might first become popular in Dallas, then in Austin, and so on.

Returning to Figure 16(a) and Figure 16(b), we observe that hashtags reach their lowest interval focus and highest interval entropy about 10-20 minutes after their peak. Rather than peaking with their most "global" footprint, hashtags instead reach this state *after* their peak. This result – that a peaking hashtag is actually more local than it ultimately will be – is seemingly counterintuitive. However, recall that in our in our examination of the cumulative distribution of focus shown in Figure 7(a), we noted that almost 50% of hashtags accumulate more than 50% of their occurrences from a single location. With this in mind, we find that hashtags receive most of their occurrences from this single location during their peak explaining the spike in interval focus and the fall in interval entropy. In effect, this single location is "championing" a hashtag. In the 10-20 minutes after this peak period, other locations adopt the hashtag, resulting in a decrease in interval focus and an increase in entropy as the hashtags becomes more global. About 30 minutes after reaching peak, focus and entropy reverse, with focus increasing and entropy decreasing as the hashtag withdraws back to its original focus location.

In essence, hashtags are spread via a single location "championing" a hashtag initially, spreading it to other locations and then continuing to propagate it after it has become popular. In [5], the authors observed a similar pattern for YouTube videos which they called the "spray-and-diffuse" pattern. Our observations over hashtags suggest that this pattern may be a fundamental property of social media spread.

6. HASHTAG-BASED SPATIAL ANALYTICS

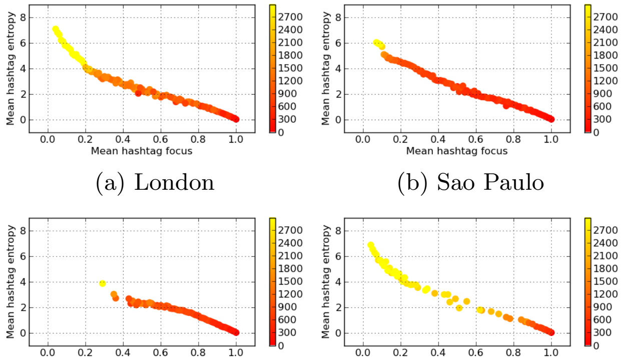
Finally, we turn our sights towards leveraging the spatio-temporal propagation of hashtags to characterize locations. Are some locations more "impactful" in terms of the hashtags that originate there, and other locations more "impressionable" in terms of hashtags they propagate? Concretely, we illustrate two techniques for characterizing locations based on hashtag spatial analytics: (i) location-based entropy-focus-spread plots; and (ii) a method for evaluating the spatial impact of locations.

6.1 Entropy-Focus-Spread Plots

In the first technique, we first assign every hashtag to its corresponding hashtag focus location. This results in every location having a set of hashtags that were focused there. Using this set of hashtags we plot the entropy versus focus for every hashtag focused on this location *plus* indicate the mean spread for every focus-entropy pair using a color gradient. To illustrate, consider the four location-based entropy-focus-spread plots in Figure 17 – one for London, Sao Paulo, Ankara, and Los Angeles. Recall that London, Sao Paulo, and Los Angeles are among the top-5 locations in terms of total hashtags, while Ankara ranks much lower.

First, we observe that locations that have high hashtag counts have a complete spectrum of hashtags on the plots. Recall that local hashtags occur on the right-bottom of such plots, while global hashtags are on the left-top. Here we see that the popular locations are the focal points (or "champions") for both local and global hashtags. Ankara, on the other hand, is the focal point for only relatively local hashtags (with high focus and low entropy).

Second, the use of spread (with high values in a lighter yellow, while lower values of spread are in red) illustrates the relative geo-spatial footprint of hashtags that have a location as its focal point. For example, although Sao Paulo has a high total number of hashtags and a high number of total locations impacted (note the hashtags with low focus and high entropy), the geospatial footprint of Sao Paulo is relatively low (note the very little yellow among these hashtags). The hashtags popular in Sao Paulo have high entropy because they are spread over several locations but all these locations are close to each other resulting in a smaller spread. Los Angeles on the other hand has a global impact; hashtags



(a) London

(b) Sao Paulo

(c) Ankara

(d) Los Angeles

Figure 17: (Color) Entropy-Focus-Spread plots for four cities. Local hashtags – with a high focus and a low entropy – are located in the bottom-right of each figure; global hashtags – with a low focus and a high entropy – are located in the top-left of each figure. The mean spread for every focus-entropy pair using a color gradient: high values in a lighter yellow, while lower values of spread are in red.

that become popular in Los Angeles tend to be popular in a larger geographical area.

6.2 Measuring Spatial Impact

The second spatial analytics technique directly evaluates the impact a location has on other locations by measuring the hashtag-based spatial impact. We define the *spatial impact* $\mathcal{I}_{l_i\to l_j}$ of location l_i on l_j as a score in the range $[-1, 1]$, such that -1 indicates l_i adopts a hashtag only after l_j has adopted it, $+1$ indicates l_j adopts a hashtag only after l_i adopts it and 0 indicates the locations are independent of each other and adopt hashtags simultaneously.

For example, consider the three cases shown in Figure 18. When hashtags are generated between a pair of locations as shown in (a) we want $\mathcal{I}_{l_1\to l_2} = 1$, when as shown in (b) we want $\mathcal{I}_{l_1\to l_2} = -1$, and when as shown in (c) we want $\mathcal{I}_{l_1\to l_2} = 0$. Let $o^h_l(t)$ represent an occurrence of hashtag h in location l at time interval t. Then, we define the preceding operator $\prec$ over two sets of occurrences $O^h_{l_i}$ and $O^h_{l_j}$ as:

$$O^h_{l_i} \prec O^h_{l_j} = \{o^h_{l_i}(t) \mid t_i < t_j \ \forall \ (o^h_{l_i}(t_1), o^h_{l_j}(t_2)) \in O^h_{l_i} \times O^h_{l_j}\}$$

which gives a set of all occurrences of h in l_1 that precede l_2 in the cartesian product of their occurrences. Similarly, we define the succeeding operator $\succ$ as:

$$O^h_{l_i} \succ O^h_{l_j} = \{o^h_{l_i}(t) \mid t_i > t_j \ \forall \ (o^h_{l_1}(t_1), o^h_{l_j}(t_2)) \in O^h_{l_i} \times O^h_{l_j}\}$$

which gives the set of all occurrences of h in l_1 that succeed l_2 in the cartesian product of their occurrences. We now define the spatial impact of location l_i on l_j as the average of hashtag specific spatial impact values, $\mathcal{I}^h_{l_i\to l_j}$, for all hashtags that occur in both the locations:

$$\mathcal{I}_{l_i\to l_j} = \frac{\sum_{h\in H_{l_i}\cup H_{l_j}} \mathcal{I}^h_{l_i\to l_j}}{|H_{l_i}\cup H_{l_j}|}$$

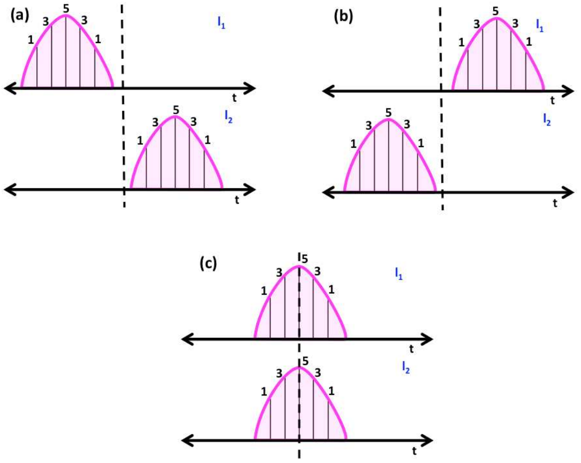


Figure 18: Example of hashtag adoption for two locations l_1 and l_2. In (a) l_1 adopts all of its hashtags before l_2. In (b) l_1 adopts all of its hashtags after l_2. In (c) l_1 and l_2 adopt hashtags simultaneously.

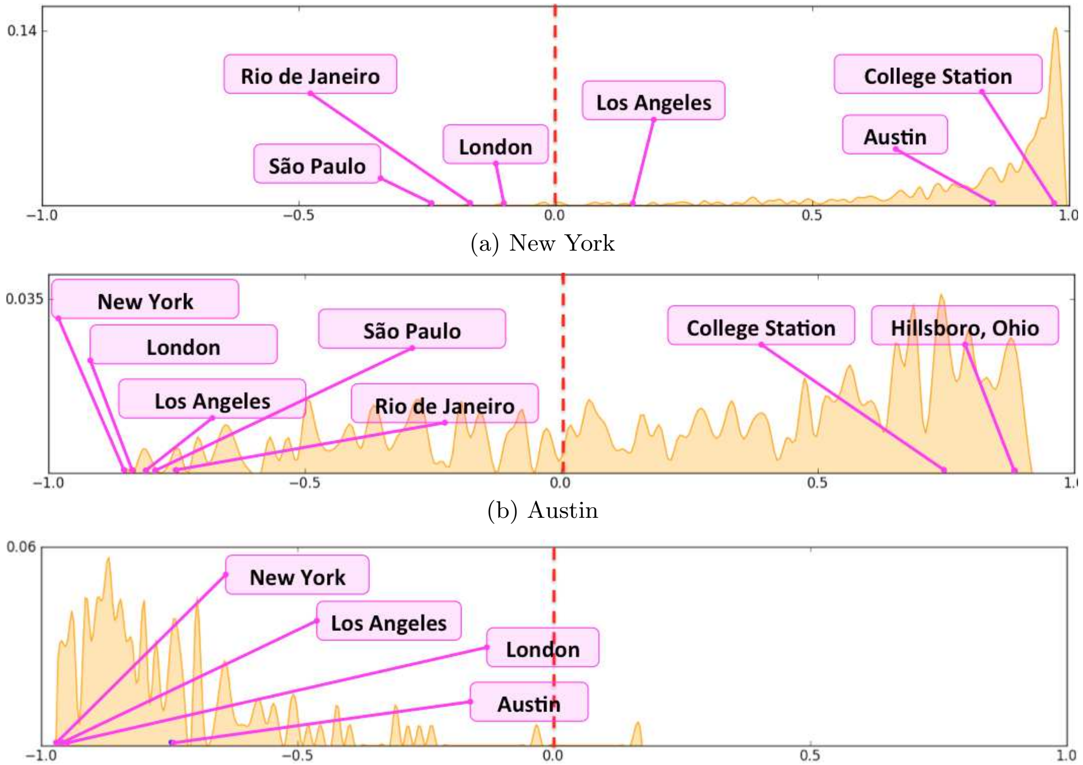
where, $\mathcal{I}^h_{l_i\to l_j}$ is defined as:

$$\mathcal{I}^h_{l_i\to l_j} = \begin{cases} \frac{|O^h_{l_i}\prec O^h_{l_j}|-|O^h_{l_i}\succ O^h_{l_j}|}{|O^h_{l_i}\times O^h_{l_j}|} & \text{if } h\in H_{l_i} \text{ and } h\in H_{l_j}\\ 1 & \text{if } h\in H_{l_i} \text{ only}\\ -1 & \text{if } h\in H_{l_j} \text{ only} \end{cases}$$

The impact is 1 if a hashtag is posted only in l_i, as l_i clearly impacts l_j in this case. For similar reasons the impact is -1 when a hashtag is posted only in l_j. To understand the case when a hashtag is observed in both the locations consider the example shown in Figure 18. In all three cases $|O^h_{l_1}| = 13$, $|O^h_{l_2}| = 13$ and $|O^h_{l_1}\times O^h_{l_2}| = 169$.

- **Case (a)**: $|O^h_{l_1}\prec O^h_{l_2}| = 169$ and $|O^h_{l_1}\succ O^h_{l_2}| = 0$. Hence, $\mathcal{I}^h_{l_i\to l_j} = \frac{169-0}{169} = 1$.
- **Case (b)**: $|O^h_{l_1}\prec O^h_{l_2}| = 0$ and $|O^h_{l_1}\succ O^h_{l_2}| = 169$. Hence, $\mathcal{I}^h_{l_i\to l_j} = \frac{0-169}{169} = -1$.
- **Case (c)**: $|O^h_{l_1}\prec O^h_{l_2}| = 62$ and $|O^h_{l_1}\succ O^h_{l_2}| = 62$. Hence, $\mathcal{I}^h_{l_i\to l_j} = \frac{62-62}{169} = 0$.

We visualize the spatial impact of a location using a *spatial impact plot*. The x-axis represents the spatial impact values and is in the range $[-1, 1]$; the y-axis shows the distribution of locations at these values. Examples of impact plots for three locations can be found in Figure 19. In every impact plot, locations on the left half of the plot are *impacting* locations and the locations on the right half of the plot are *impacted* locations. Hence, plots for famous and large locations are generally right-heavy as they impact many locations. Plots for small locations are mostly left-heavy as they are impacted by many locations. For example, the impact plot for New York is right heavy since New York is an "early adopter" with a high spatial impact on other locations. Interestingly, New York is actually impacted by both Sao Paulo and Rio de Janeiro, since Portuguese hashtags tend to flow from Brazil to Portuguese-speaking neighborhoods in New York, whereas hashtags from New York are less likely to flow to Brazil. College Station (home to Texas A&M) is fairly small, with a left-heavy distribution, indicat-



(a) New York

(b) Austin

(c) College Station

Figure 19: Spatial impact plots for three locations. Locations to the left of the origin are "early adopters" relative to the baseline location. New York has a high impact, with almost all cities to the right of its origin. College Station, on the other hand, is low impact since it only adopts hashtags after almost all other cities.

ing that it is a "late adopter". Austin, on the other hand, has a balanced spatial impact, being both impacted by many locations and impacting many other locations.

7. CONCLUSION

In this paper, we have analyzed the spatio-temporal dynamics of social media propagation through a study of 2 billion geo-tagged Tweets. Our study has consisted of three key parts: (i) a study of the global footprint of hashtags and an exploration of the spatial constraints on hashtag adoption; (ii) a study of three spatial properties of hashtag propagation – focus, entropy, and spread – and an examination of the spatial propagation of hashtags using these properties; and (iii) two spatial analytics techniques for characterizing the relative impact of locations. We have found that hashtags are a global phenomenon, with locations all across the world. But the physical distance between locations is a strong constraint on the adoption of hashtags, both in terms of the hashtags shared between locations and in the timing of when these hashtags are adopted. We have also found that hashtags are mostly a local phenomenon with long-tailed life spans, but follow a "spray-and-diffuse" pattern [5] where initially a small number of locations "champion" a hashtag, make it popular, and the spread it to other locations. We have found both spatial and temporal locality as most hashtags spread over small geographical areas but at high speeds. The purpose of a hashtag and its global awareness determines how fast it will reach its peak. A hashtag representing a globally known event reaches its peak much faster than either locally-known events or hashtags spread purely within the network (e.g., #ifyouknowmeyouknow). Based on spatial and temporal categories we classified hashtags into different categories. In our continuing work we are interested in hashtag category specific analysis. We want to study how the temporal characteristics of hashtags may differ depending upon their spatial categories.

8. ACKNOWLEDGMENTS

This work was supported in part by NSF grant IIS-1149383 and a Google Research Award. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

9. REFERENCES

[1] Universal transverse mercator coordinate system, November 2012.

[2] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *Proceeding of the 17th international conference on World Wide Web*, pages 357–366. ACM, 2008.

[3] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[4] C. Bauckhage. Insights into internet memes. *Proc. ICWSM2011*, pages 42–49, 2011.

[5] 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.

[6] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[7] Z. Cheng, J. Caverlee, K. Lee, and D. Sui. Exploring millions of footprints in location sharing services. *AAAI ICWSM*, 2011.

[8] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. Gonçalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 58–65, 2011.

[9] N. Dalvi, R. Kumar, and B. Pang. Object matching in tweets with spatial models. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 43–52. ACM, 2012.

[10] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 241–249. Association for Computational Linguistics, 2010.

[11] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. ., 2000.

[12] Foursquare. About foursquare, April 2013.

[13] B. A. Huberman, D. M. Romero, and F. Wu. Social Networks that Matter: Twitter Under the Microscope. *Social Science Research Network Working Paper Series*, Dec. 2008.

[14] K. Y. Kamath, J. Caverlee, Z. Cheng, and D. Z. Sui. Spatial influence vs. community influence: modeling the global spread of social media. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 962–971, New York, NY, USA, 2012. ACM.

[15] 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.

[16] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[17] 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.

[18] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 422–429. ACM, 2011.

[19] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. *ICWSM'11*, 2011.

[20] D. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

[21] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of ICWSM*, 11:329–336, 2011.

[22] O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 643–652. ACM, 2012.

[23] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 599–608. IEEE, 2010.