Spatial Influence vs. Community Influence: Modeling the Global Spread of Social Media

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

Daniel Z. Sui
Department of Geography
Ohio State University
Columbus, OH 43210
sui.10@osu.edu

ABSTRACT

In this paper we seek to understand and model the global spread of social media. How does social media spread from location to location across the globe? Can we model this spread and predict where social media will be popular in the future? Toward answering these questions, we develop a probabilistic model that synthesizes two conflicting hypotheses about the nature of online information spread: (i) the spatial influence model, which asserts that social media spreads to locations that are close by; and (ii) the community affinity influence model, which asserts that social media spreads between locations that are culturally connected, even if they are distant. Based on the geospatial footprint of 755 million geo-tagged hashtags spread through Twitter, we evaluate these models at predicting locations that will adopt hashtags in the future. We find that distance is the single most important explanation of future hashtag adoption since hashtags are fundamentally local. We also find that community affinities (like culture, language, and common interests) enhance the quality of purely spatial models, indicating the necessity of incorporating non-spatial features into models of global social media spread.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

social media, information diffusion models, virtual communities

1. INTRODUCTION

Users generate and consume a great deal of content on the Internet every day in the form of videos, blogs, tweets, and so on. YouTube, for example, streams more than 4 billion videos everyday, with 60 hours of new content uploaded every minute [21]. As users consume and share this content, some of it tends to gain traction and become popular resulting in viral videos, trending hashtags, popular blogs, and so forth. These phenomena have attracted a considerable amount of recent research to study the dynamics of the adoption of social media [3, 12, 16, 18, 22].

Of particular importance is the geospatial spread of social media. For example, how did videos captured on smartphones during the Arab Spring spread across the globe? Are there key locations that promoted the spread of these videos? As the Arab Spring has become increasingly part of the US's social conscious, do we see key US locations impacting the propagation of videos today? Answering these questions is extremely challenging, and so as a beginning step we study in this paper the dynamics of social media adoption across geographical locations. Concretely, we formalize the problem of predicting the global spread of social media as the *location subset selection problem*. That is, as a particular item (e.g., video, image) begins to propagate can we predict the locations where it will soon become popular? For example, observing a video which is gaining traction in Qatar, can we predict locations in Europe where the video is soon going to become popular?

Previous work in the area of information (content) diffusion and influence propagation have tended to focus on the pathways of diffusion through social and information networks, e.g., [11, 13, 14, 15, 17, 26]. Complementary to these efforts, we focus on the geospatial connections that impact the spread of social media, and so we abstract from the interaction network layer to consider fine-grained locations and their connections to other locations. Towards modeling the global spread of social media, we develop a probabilistic model that synthesizes two conflicting hypotheses about the nature of online information spread:

- **Distance matters**. As encapsulated by Tobler's first law of geography [24] which asserts that all things being equal, closer places are more alike, whereas distance places are more unalike. In the context of social media spread, Tobler's first law of geography would suggest that locations that are close to each other should be more likely to adopt similar online behaviors (e.g., viewing a YouTube video, posting the same hashtag).
- **"Distance is dead"** [5]. The second hypothesis claims that since online interactions are freed from geospatial constraints, mere proximity is no guarantee toward adopting similar online behavior. In this setting, long-distance links formed through common online community may be more predictive. For example, tech communities in Austin, San Francisco, and Seattle may be tightly linked through their common interest in similar YouTube videos, whereas more geographically close locations may share little in common.

Based on the first hypothesis, we develop the *spatial influence model*, which asserts that the adoption of a particular user activity in a nearby location has a stronger influence on a target location than

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

whether that same activity was adopted at a more distant location. In other words, distance matters. Based on the second hypothesis, we develop the *community affinity influence model*, which asserts that locations that share a similar community affinity, regardless of distance from each other, are more likely to influence one another. While there are many ways to measure community affinity, we propose two methods: (i) the first considers communities to be close to each other if they share similar activities regardless of *when* they adopt these activities, for example tech communities in Austin and San Francisco reading similar articles on thehackernews.com; and (ii) the second considers communities close to each other if they tend to adopt similar activities in sync, like a video becoming popular in New York and Boston around the same time. Note that both the spatial influence model and the community affinity influence model are developed completely orthogonal to the underlying social network and are based solely on the geospatial distribution of user activities, meaning that estimating flows of influence from one person to another are not necessary. We test these models in the context of the geospatial footprint of 755 million geo-tagged hashtags spread through Twitter. We find that while the spatial influence model has a higher impact than the community affinity influence model in predicting the spread, its combination with community affinity influence model gives the best performance, suggesting that both distance and community are key contributors to social media spread.

The rest of the paper is organized as follows. We start by describing related works in Section 2. In Section 3, we describe our dataset and measure geo-spatial properties of social media propagation. In Section 4, we formally define the location subset selection problem and present the spatial influence and community affinity models. Finally, in Section 5, we define the metrics to compare these models and evaluate the performance of these models before concluding in Section 6.

2. RELATED WORK

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

Information Diffusion on Twitter: There have been several papers studying the general properties of Twitter as a social network and in analyzing information diffusion over this network [12, 16, 17, 26]. 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 [22] 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 been to study the property of hashtag creation, use, and dissemination in [7]. In related research, approaches based on linear regression have been used to predict the popularity of hashtags in a given time frame in [25]. Because of the semantic nature of hashtags and the variety of ways in which it is used to convey information about a tweet, there have been some papers which have used hashtags to solve problems like sentiment detection [9], topic tracking on twitter streams [19], 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 [23, 20]. Some of the earliest research related to geo-spatial analysis of web content were based on mining geography specific content for search engines [10]. More recently in [1] 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 [8]. Similar analysis to infer user's location on Facebook based on their social network has been studied in [2]. A recent paper dealt with the spatial analysis of Youtube videos [4] . In this work the authors were able to observe the highly local nature of videos based on the propagation patterns of Youtube videos.

3. MEASURING THE GEOSPATIAL PROPERTIES OF SOCIAL MEDIA

In this section we first present notation for measuring social media spread with an eye toward developing models of this spread. Then we highlight the experimental setting – Twitter-based hashtags – and examine the geospatial properties of hashtag spread. Our goal is to study questions like: Does distance impact whether social media (hashtags, in this case) is shared between two locations? Does distance impact the timing of hashtag adoption? How predictable is the spread of a hashtag over a geographic area? Do early observations indicate whether a hashtag will spread compactly or be widely diffused over a large spatial area?

3.1 Preliminaries

Let M be the set of user activities of interest – for example, an activity could correspond to a click on a web link, a view of a Web video, sharing of a link on Facebook, posting a particular hashtag on Twitter, and so on. Suppose we have divided the globe into a set of distinct locations L (say by overlaying a mesh dividing the globe into squares of 0.001 degrees latitude by 0.001 degree longitude). Every activity is associated with some subset of locations in which the activity has been observed. For example, based on the IP address, a view of a Web video can be traced back to an approximate latitude and longitude. Similarly, many social media services and smartphones support GPS-enabled tagging of user activities. By discretizing time into regular intervals (say, into 5 second increments), we can express the set of occurrences of an activity $m \in M$ in a particular location $l \in L$ at time t as $o_l^m(t)$. For example, o_l^m may represent 10 clicks of a Web video m in the past minute, where each click originates in a particular neighborhood l.

Now, suppose we have observed all occurrences of an activity up to some critical time t_s. Then we can define the set of **observed occurrences** (O_l^m) of m at a single location l as:

$$O_l^m = \bigcup_{t=0}^{t_s} o_l^m(t) \tag{1}$$

and the **total observed occurrences** set O^m across all locations in L as:

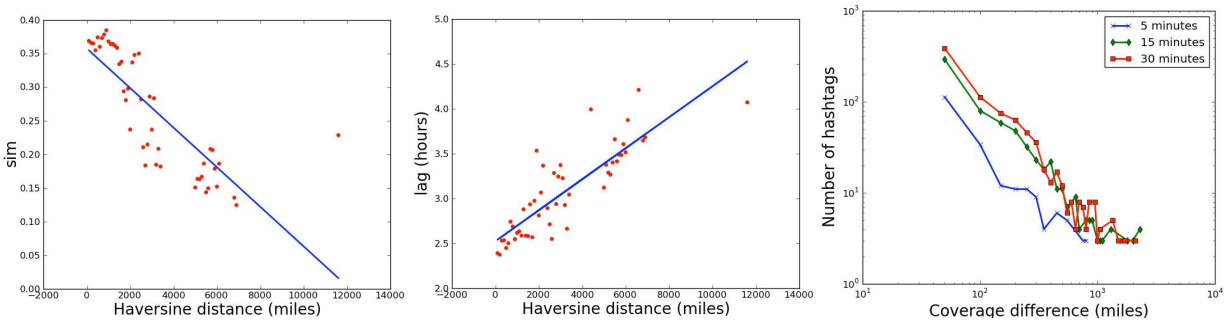
$$O^m = \bigcup_{l \in L} O_l^m$$

We denote the set of unique hashtags observed in l as M_l.

3.2 Experimental Setting: Hashtags

To measure the geospatial properties of social media, we focus our attention on one type of globally observed user activity – the posting of hashtags on Twitter. Twitter hashtags are prefixed with a # and mostly serve as tags to the corresponding tweet. Users tag their tweets for different purposes. For example, some are event driven like #ripstevejobs, and #fukushima, while some are mostly for fun like #bestsportsrivalry and #ifyouknowmeyouknow.

We collected a sample of around ~755 million geo-tagged tweets containing ~10 million unique hashtags from Twitter using the Twitter Streaming API from February 1 to November 30, 2011. Each tweet in this sample is tagged with a latitude and longitude indicating the location of the user at the time of the posting. All $<$ `hashtag`, `time`, `latitude`, `longitude` $>$ tuples corresponding to a particular hashtag are considered as a single activity of interest. Together all hashtags give us the set of all activities M.



(a) Hashtag sharing versus distance (b) Hashtag adoption lag versus distance (c) Early indication of coverage.

Figure 1: Geospatial properties of hashtags. (a) shows correlation between location similarity and distance. We see that similarity between location decreases with increasing distance. (b) shown correlation between hashtag adoption lag and distance. We see that adoption lag increases with increasing distance. (c) comparison between early and late coverage for call hashtags. A power law indicates that most hashtags have a small difference between early and late coverage values.

We round latitudes and longitudes to their nearest tenth values, which overlays a mesh dividing the globe into locations (L). To avoid sparsely represented hashtags, we consider only hashtags with at least 5 occurrences in a location and consider only hashtags with at least 250 total occurrences across all locations. Since some hashtags may have begun their Twitter life before the first day of our sample (February 1) while others may have continued on after the last day (November 30), we consider both February and November as buffer months. Hence, we capture the full lifecycle of hashtags starting on or after March 1 and ending by October 31, which focuses our study to hashtags which have both their birth and death within the time of study (and as a result, removes cyclical hashtags like "#ff" and "#nofollow"). We additionally divide the set of all hashtags into two sets: a training set based on hashtags from March to August; and a test set based on September to October. Hashtags that start in training but continue into test are ignored. In this way, the training set contains 1466 complete hashtag propagations and the test set contains 515.

3.3 Geospatial Properties of Hashtags

Toward informing the development of models of social media spread, we study three geospatial properties of hashtags: (i) sharing versus distance, (ii) adoption lag versus distance, and (iii) the predictability of spread.

Hashtag Sharing versus Distance: We first seek to understand the relationship of the distance between locations on the commonality of hashtags adopted in locations. Do we find that distance has no impact on whether a hashtag is shared between two locations? We define the distance between two locations using the Haversine distance, which is commonly used to measure the distance between locations based on the spherical shape of the Earth (as compared to Euclidian distance)[1]. In essence, the Haversine maps from latitude-longitude pairs to distance: $\mathcal{D} : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$. $H : L \times L \rightarrow \mathbb{R}_{\geq 0}$.

Given two locations, we measure their hashtag "similarity" using the Jaccard coefficient between the sets of hashtags observed at each location:

$$sim_{hashtag}(l_1, l_2) = \frac{M_{l_1} \cap M_{l_2}}{M_{l_1} \cup M_{l_2}}$$

where recall M_l is the 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 1(a). We see a strong correlation ($\rho = -0.8$), suggesting that the more distant two locations are, the less alike they are. We also note that, though the similarities are high for most location pairs that are close to each other, there are some location pairs (above the blue line) where this doesn't hold true. Presumably, these outliers are linked by some other factors (language, culture), which we shall explore in the community affinity model shortly.

Hashtag Adoption Lag versus Distance: We additionally can measure the lag between two locations by measuring how close in time did the two locations adopt the same hashtag. Locations that adopt a common hashtag at the same time are more similar (and have a smaller lag) than are two locations that are farther apart in time (with a greater lag). Letting M_l be the set of unique hashtags observed in l and t_l^m be the time of first occurrence of m at l, we can define the hashtag adoption lag of two locations as:

$$lag_{adoption}(l_1, l_2) = \frac{1}{|M_{l_1} \cap M_{l_2}|} \sum_{m \in M_{l_1} \cap M_{l_2}} |t_{l_1}^m - t_{l_2}^m|$$

where the adoption lag measures the mean temporal lag between two locations for hashtags that occur in both the locations. A lower value for this measure indicates that common hashtags appear to reach both the locations around same time. We see in Figure 1(b) a positive correlation ($\rho = 0.86$), suggesting that locations that are close in spatial distance tend also to be close in temporal distance (e.g., they adopt hashtags at approximately the same time). Locations that are more spatially distant tend to adopt hashtags at much greater lags with respect to each other. As in the case of hashtag sharing, we see many location pairs having low lags despite being quite distant from each other, suggesting some other mechanism is at work.

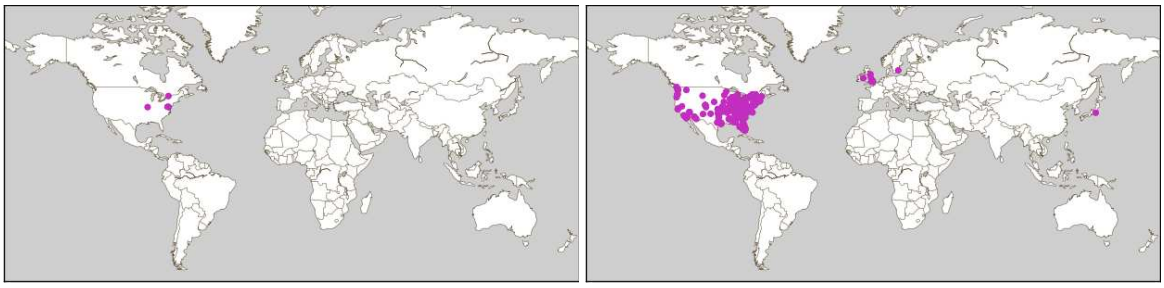
Predictability of Spread: Finally, we measure the predictability of the "spread" of hashtag over a geographic area through its *coverage*. Coverage measures the mean Haversine distance for all occurrences of a hashtag from its geographic midpoint:

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

where we define the geographic midpoint[2] for a set of occurrences as a function $G : O \rightarrow \mathbb{R}^2_{\geq 0}$, where the first dimension is the

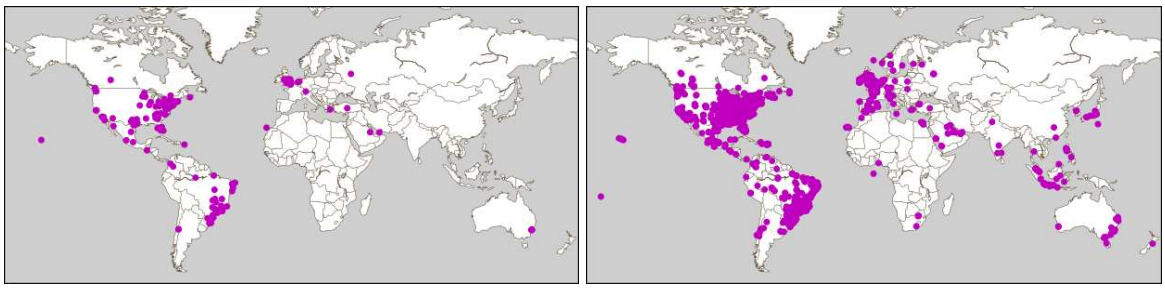
[1] For a fuller treatment, we refer the interested reader to http://en.wikipedia.org/wiki/Haversine_formula

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



(a) (b)

Figure 2: #cnndebate after 5 minutes (left) and 2 hours (right)



(a) (b)

Figure 3: #ripstevejobs after 5 minutes (left) and 2 hours (right)

latitude and the second is the longitude of the midpoint. The calculation of geographic midpoint 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 hashtag localized to a specific areas has a small coverage, while a universal hashtag has a larger coverage. To illustrate, consider the two hashtags #cnndebate and #ripstevejobs. Figure 2(b) shows the propagation of #cnndebate – corresponding to the Republican Presidential debate – after 2 hours. We see that the hashtag is mostly local to the United States and has a coverage of 743.32 miles. In contrast, Figure 3(b) shows the propagation of #ripstevejobs after 2 hours, resulting in a coverage of 3120.96 miles, indicating a global footprint.

To understand the predictability of spread, we measure the distribution of differences between the coverage for hashtags after they have completely propagated and coverage after the hashtag has propagated for a smaller time interval. For three initial periods – of 5 minutes, 15 minutes, and 30 minutes – we plot the difference between the coverage at this early time of a hashtag's propagation and the coverage after the completion of the hashtag's entire lifespan. We observe in Figure 1(c) that most hashtags have a small coverage difference, indicating that the final coverage of hashtag propagations can be accurately estimated early in its lifecycle. And the predictability of coverage increases as the length of the initial period increases (from 5 to 30 minutes); that is, as more evidence is accumulated over the beginning stages of a hashtag, the final coverage differs by less.

Continuing the example of #cnndebate and #ripstevejobs, we see in Figure 2 and Figure 3 that occurrences observed early in a hashtag's lifecycle (in this case, after just 5 minutes) are good indicators of later occurrences (in this case, measured after 120 minutes).

Based on these three geospatial properties, we observe:

1. In most cases, pairs of locations that are close to each other tend to share common hashtags and adopt them around the same time, compared with locations that are distant.

2. Many distant location pairs, though, exhibit similar patterns of "closeness" in that they share hashtags and have a low hashtag adoption lag, suggesting some additional factor is "bending space" to link the two locations.

3. Finally, the spread predictability analysis suggests that early occurrences of a hashtag are good indicators of the relative coverage of a hashtag's future spread (either compact or widely diffuse).

4. MODELING HASHTAG SPREAD

Based on these observations, we next turn to the challenge of developing models of hashtag spread. Specifically we develop and evaluate the *spatial influence model* – in which nearby locations strongly influence hashtag adoption – and the *community influence model* – in which "similar", though perhaps distant, locations strongly influence hashtag adoption. The intuition behind both approaches is that locations influence each other, and that the future spread of a hashtag is guided by this mutual influence.

4.1 Problem Setting

To formalize the development of such hashtag spread models and to provide an experimental grounding for evaluating the quality of such models, we focus on the problem of selecting future locations that will adopt a hashtag based on the partial evidence of the hashtag's propagation up until that time. We call this the *location subset selection problem*. That is, as a particular social media begins to propagate can we predict the locations where it will soon arrive and become popular? For example, observing a video which is gaining traction in Qatar, can we predict locations in Europe where the video is soon going to become popular? The models developed for tackling this problem are an important and necessary step for

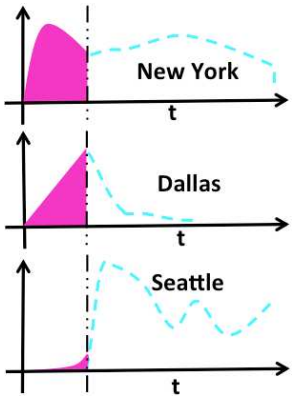


Figure 4: Based on the observed postings of a hashtag up to some time t_s (the vertical dotted line), can we predict which locations will post the most hashtags in the future?

supporting content localization, geo-advertising, fraud detection, and other social media analytics. It is particularly important that such models robustly predict the spread of social media while it is still developing (e.g., a video is going viral, a meme is becoming increasingly popular).

Recall the **total observed occurrences** set O^m across all locations in L ($O^m = \bigcup_{l \in L} O_l^m$) introduced in Section 3.1. In practice, these observed activities will vary by location. Early adopting locations may encompass many postings of a hashtag (or views of a Web video, ...), while later adopting locations will have few or no postings of a hashtag (or views of a video, ...), especially in the early moments of a hashtag's rise to popularity. Based on this state up to some time t_s, can we select some subset of locations $S \in L$ such that these locations are likely to observe many occurrences of the user activity.

For example, consider the three locations – New York, Dallas and Seattle – shown in Figure 4 and suppose a particular hashtag has been posted from each location. Based on the observed hashtag postings up to some time t_s (the vertical dotted line), can we predict which locations will post the most hashtags in the future? Toward this goal, we can express the occurrences of the activity *after the critical time* t_s as the unknown future set of **unobserved occurrences**:

$$U_l^m = \bigcup_{t=t_s+1}^{\infty} o_l^m(t) \tag{2}$$

where U_l^m is the set of occurrences of m observed in location l after time t_s. We can additionally express the **total unobserved occurrences** set U^m across all locations in L as:

$$U^m = \bigcup_{l \in L} U_l^m$$

Together, the total occurrences of an activity throughout its lifetime is $O^m \cup U^m$. Now, suppose for some subset of locations $S \subseteq L$, we measure the count of the total unobserved occurrences of an activity in this subset as U_S^m:

$$U_S^m = \sum_{l \in S} |U_l^m|$$

We can then formulate the task of selecting the best k locations at some critical time t_s as the **location subset selection problem**:

DEFINITION 4.1. ***(Location Subset Selection Problem)**: Given an integer k, the location subset selection problem for a user activity m at time t_s is the problem of predicting top-k locations which will have the highest number of unobserved occurrences for m.*

$$\mathcal{M}(m, L) = S_{t_s}^m = \underset{\{S \subseteq L \mid |S|=k\}}{\arg\max} U_S^m$$

where, $\mathcal{M} : M \times L^{|L|} \rightarrow L^k$, defined as subset selection model, takes a user activity and the set of all locations as input and returns a subset of locations of cardinality k.

The challenge for identifying the best choice of locations $S_{t_s}^m$ at time t_s is difficult because the future occurrences set for all locations, U^m, is available only after the complete evolution of the activity of interest. Hence, we must predict which locations are the best. Of course, determining the best choice of locations is simpler the longer the decision point is delayed (since many bursting and trending phenomenon will have run their course, saturating its locations), but of less value. The question is whether the best set of locations $S_{t_s}^m$ can be identified for some time t_s close to the activity's first observed occurrence.

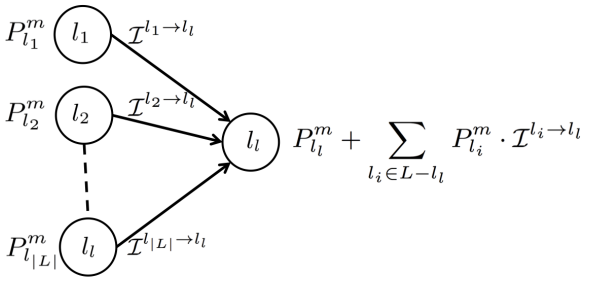


Figure 5: General spatial influence model.

4.2 Modeling Spread: Spatial Influence vs. Community Influence

With the problem statement in mind as well as our observations of the geospatial spread of hashtags, we now propose location influence based models for geo-spatial spread. The intuition behind our approach is that locations influence each other. And given a hashtag distribution, the future propagation of this hashtag is guided by this mutual influence between locations. The influence exerted by a location on another could be based either on proximity between locations or on the culture, language, and common interests shared by these locations. We measure this influence using an influence metric $\mathcal{I}^{l_i \rightarrow l_j}$ which has a range of $[0, 1]$ and represents the influence location l_i has on l_j such that the higher the value of this metric, then the greater is the influence exerted by l_i on l_j.

So given a hashtag m, the spread model for an influence metric $\mathcal{I}^{l_i \rightarrow l_j}$ is defined as:

$$\mathcal{M}_{\text{Spread}}(m, L) = \underset{\{S \subseteq L \mid |S|=k\}}{\arg\max} \sum_{l \in S} \left(P_l^m + \sum_{l_i \in L - l} P_{l_i}^m \cdot \mathcal{I}^{l_i \rightarrow l} \right)$$

where, $P_l^m = \frac{|O_l^m|}{|O^m|}$ is the probability of observing user activity m in l, estimated based on m's propagation until t_s and the expression within the parenthesis calculates the total effective influence exerted at this location to generate m. This concept is shown in Figure 5, where the location l_l gets influenced by all the locations and the effective influence on it is calculated as shown above. The spread model relies on the third observation that early occurrences of a hashtag are good predictors of future coverage. Hence, in this expression we use the probability of observing m in l to modify l's influence while calculating the effective influence. In this way the

spread model, $\mathcal{M}_{\text{Spread}}$, selects a subset of the most influenced locations with the belief that this influence will make these locations adopt hashtags in future.

Using the spread model as framework, we now describe two general approaches – the spatial influence model and the community affinity model – that build on the observations made in Section 3.

4.2.1 Spatial Influence Model

The spatial influence model is based on our first observation in Section 3.3 that tells us that distance between locations influences what hashtags are shared and when they are shared. So, we define the spatial influence metric, $\mathcal{I}_{\text{Spatial}}^{l_j \to l_i}$, as:

$$\mathcal{I}_{\text{Spatial}}^{l_j \to l_i} = \frac{\alpha^{-H(l_i,l_j)}}{\sum_{l_i \in L} \alpha^{-H(l_i,l)}}$$

where, the numerator exponentially decays l_i's influence on l as a function of their Haversine distance and the denominator normalizes this influence so that $\sum_{l \in L} \mathcal{I}_{\text{spatial}}^{l \to l_i} = 1.0$. The parameter α controls the rate of influence decay. A higher value for α decreases influence from a point at a higher rate and a lower value for alpha (> 1.0) decreases influence at a lower rate. Using the this influence metric we define the spatial influence model as:

$$\mathcal{M}_{\text{Spatial}}(m, L) = \underset{\{S \subseteq L \mid |S|=k\}}{\arg\max} \sum_{l \in S} \left(P_l^m + \sum_{l_i \in L-l} P_{l_i}^m \cdot \mathcal{I}_{\text{Spatial}}^{l_i \to l} \right) \tag{3}$$

To illustrate, consider an example of a hashtag that occurs only in Houston. Now given an option between Austin and San Francisco, the model as defined in (3) picks Austin since it is much closer to Houston than San Francisco.

A real world example of modeling propagations using the spatial influence model for the hashtag #ripstevejobs is shown in Figure 6. We predicted the future distribution of this hashtag using the spatial influence model based solely on its initial (first 5 minutes) distribution. The comparison between the predicted and actual distribution is shown in Figure 6(a) and Figure 6(b) respectively. We observe that the relative distribution (indicated by color) and its values (indicated by scale) are very close to each other.

4.2.2 Community Affinity Influence Model

Of course, distance is not the only factor that impacts the spread of a hashtag, as we observed in Section 3.3 (second observation). Hence, we now propose the community affinity influence models for capturing non-distance links between locations like culture, language, and common community interest. Concretely, we define two influence metrics to model community affinity based on their common usage of hashtags.

- **Transmitting Influence**: Using temporal proximity, we observe that if a hashtag is observed at a particular location, then it will soon be observed in other related locations as well. To model the degree to which a location can impact other locations temporally, we define the transmitting score, $\mathcal{T}$, as:

$$\mathcal{T}_{l_j \to l_i} = \frac{|\{m \mid t_{l_j}^m > t_{l_i}^m \quad \forall m \in M_{l_i} \cap M_{l_j}\}|}{|M_{l_i}|}$$

where, the numerator is the number of hashtags that occurred in l_1 before l_2. So, when all hashtags occurring in l_1 have occurred in l_2 and all before occurring in l_2, the transmitting score for l_1 transmitting a hashtag to l_2 - $P_t(l_2|l_1) = 1.0$. Using this we define the transmitting influence as:

$$\mathcal{I}_{\text{Trans.}}^{l_j \to l_i} = \frac{\mathcal{T}_{l_j \to l_i}}{\sum_{l \in L} \mathcal{T}_{l \to l_i}} \tag{4}$$

A value for $\mathcal{I}_{\text{Trans.}}^{l_j \to l_i}$ is in the range $[0, 1]$, with 0 indicating l_j doesn't transmit anything to l_i and 1.0 indicating l_j is the only location influencing l_i and it gets all of its hashtags after l_j.

- **Sharing Influence**: Similar to transmitting influence, we use content-related proximity to model the impact a location can have on nearby locations, using the sharing score:

$$\mathcal{S}_{l_j \to l_i} = \frac{|M_{l_i} \cap M_{l_j}|}{|M_{l_i}|}$$

This function measures the probability that l_i observes the same hashtags as l_j. Using this we define the sharing influence as:

$$\mathcal{I}_{\text{Share}}^{l_j \to l_i} = \frac{\mathcal{S}_{l_j \to l_i}}{\sum_{l \in L} \mathcal{S}_{l \to l_i}} \tag{5}$$

A value for $\mathcal{I}_{\text{Share}}^{l_j \to l_i}$ is in the range $[0, 1]$, with 0 indicating l_j doesn't share anything with l_i and 1.0 indicating l_j is the only location the influencing l_i and all hashtags that have occurred in l_i have occurred in l_j.

As in the case of the spatial influence model, we can use these two community affinity influence metrics to generate a model as:

$$\mathcal{M}_{\text{Trans.}}(m, L) = \underset{\{S \subseteq L \mid |S|=k\}}{\arg\max} \sum_{l \in S} \left(P_l^m + \sum_{l_i \in L-l} P_{l_i}^m \cdot \mathcal{I}_{\text{Trans.}}^{l_i \to l} \right)$$

which models spread using transmitting influence, and,

$$\mathcal{M}_{\text{Share}}(m, L) = \underset{\{S \subseteq L \mid |S|=k\}}{\arg\max} \sum_{l \in S} \left(P_l^m + \sum_{l_i \in L-l} P_{l_i}^m \cdot \mathcal{I}_{\text{Share}}^{l_i \to l} \right)$$

which models spread using sharing influence.

To give a bit more insight into these two models, we constructed two directed graphs over the hashtag dataset – one graph for transmitting and other for sharing influence – with locations as nodes and the influence scores calculated using these functions as edge weights. In this graph, a cluster represents a collection of nodes (locations) that are close to each other, where closeness is defined either temporally (via transmitting influence) or based on content (via sharing influence). If the functions models location relationships correctly, then nodes that are close to each other in terms of distance should be in the same cluster (observation 1) and, nodes that are culturally similar should be the same cluster (observation 2). The results from this experiments are shown in Figure 7(a) and Figure 7(b), where every cluster is represented with a different color. In both these figures we can verify the two observations. Most locations which are close to each other are in the same cluster and some locations that are culturally similar, like the locations between English speaking parts of Western Europe and United States, and French speaking parts of Brazil and France, are in the same cluster.

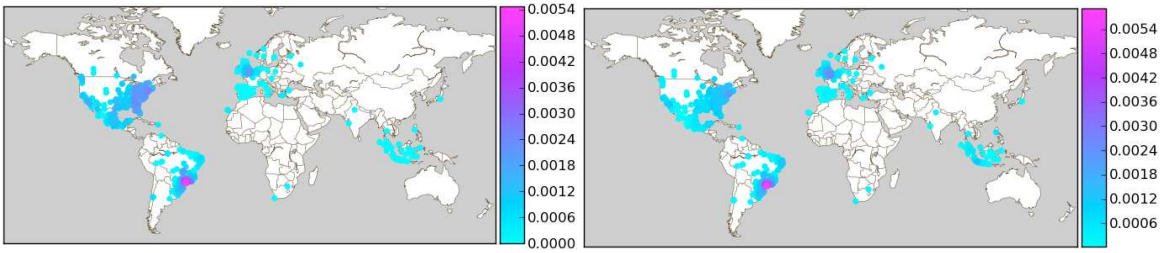
4.2.3 Combining the Two Models

We can also combine the spatial and community affinity models by first defining an effective influence score:

$$Score(l) = P_l^m + \sum_{l_i \in L-l} P_{l_i}^m \cdot (\beta \cdot \mathcal{I}_{\text{Spatial}}^{l_i \to l} + (1 - \beta) \cdot \mathcal{I}_{\text{Transmit}}^{l_i \to l}) \tag{6}$$

where, β decides the weight assigned to each model and then using to model spread as:

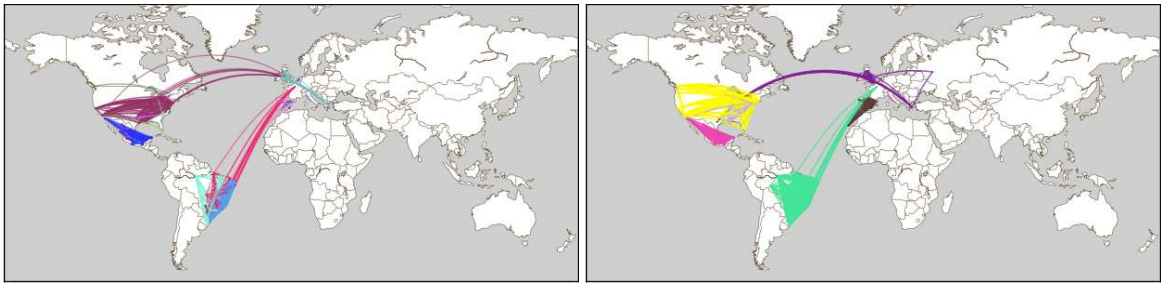
$$\mathcal{M}_{\text{Spatial + Transmit.}}(m, L) = \underset{\{S \subseteq L \mid |S|=k\}}{\arg\max} \sum_{l \in S} Score(l)$$



(a) Predicted (estimated using spatial influence model after 5 minutes)

(b) Actual (real distribution after 2 hours of propagation)

Figure 6: Example of using spatial influence model for the hashtag #ripstevejobs



(a) Transmitting Probability

(b) Sharing Probability

Figure 7: Clusters of related locations based on the transmitting and sharing probability functions.

We can define a similar model using sharing influence instead of transmitting influence as done above.

5. EXPERIMENTS

In this section, we compare the quality of the proposed location selection approaches against three baseline approaches. We introduce metrics for measuring the quality of a selection approach, investigate the proposed approaches with respect to these quality metrics and identify the best approach to solve the location selection problem.

5.1 Baseline Approaches

In addition to the three geo-spatial approaches introduced in this paper, we also consider three alternatives:

Random Selection: In this simplest approach, we randomly select k locations as the target subset, from the set of locations where the hashtag has occurred prior to t_s. The main drawback of this approach is that locations are selected without regard for the number of hashtags observed. In addition, since the target subset is selected based solely on a hashtag's propagation, the locations outside this set will never be selected. Hence, if the hashtag has occurred in fewer than k locations, then the target subset contains always fewer than k locations.

Greedy Selection: A natural improvement over random selection is a greedy approach, in which locations are selected based on the notion that a hashtag is going to continue to be used in locations where it is currently popular. Concretely, the greedy approach ranks locations based on the observed occurrence count of the hashtag: $|O_l^m|$. The intuition is that a hashtag that is popular in New York at location subset selection time is going to stay popular in the future as well. As in the random selection approach, it is possible that a hashtag might not have propagated to k locations, in which case we pick all the locations resulting in a subset with cardinality lesser than k.

Selection Based on Linear Regression: In this approach, we solve the location subset selection problem using a linear regression model. The idea behind this approach is to learn a model that can predict the unobserved occurrences for a hashtag given occurrences observed until the location subset selection time. Let M be the training hashtag set described in Section 3.2. Using M we first define the matrix X for observed occurrences as shown below:

$$X_i = \begin{pmatrix} 1 & \frac{|O_1^i|}{|O^i|} & \frac{|O_2^i|}{|O^i|} & \cdots & \frac{|O_{|L|}^i|}{|O^i|} \end{pmatrix} \quad \forall i \in [1, |M|]$$

$$X = \left(\frac{|O_j^i|}{|O^i|}\right)_{|M|\times 1+|L|} = \begin{pmatrix} X_1 & X_2 & \cdots & X_{|M|} \end{pmatrix}^T$$

where, each row in this matrix corresponds to a hashtag in the training hashtag set. Similar to X, we define the unobserved matrix Y using unobserved occurrences.

$$Y_j = \begin{pmatrix} \frac{|U_j^1|}{|U^1|} & \frac{|U_j^2|}{|U^2|} & \cdots & \frac{|U_j^{|M|}|}{|U^{|M|}|} \end{pmatrix}^T \quad \forall j \in [1, |M|]$$

$$Y = \left(\frac{|U_j^i|}{|U^i|}\right)_{|M|\times|L|} = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_{|L|} \end{pmatrix}$$

Using these matrices, we define Y as a linear function of X, $Y = X\beta + \mathcal{E}$, where, β is the $(|L| \times |L|)$ parameters matrix and $\mathcal{E}$ is the $(|L| \times |M|)$ matrix of error terms. Every column, β_l, in β models the relationships of a location l with the rest of locations and can be estimated by linear regression using the equation, $Y_l = X\beta_l + \mathcal{E}_l$, where $\mathcal{E}_l$ is the error column for l, in $\mathcal{E}$. We for a new hashtag m we can determine the top-k locations using:

$$\mathcal{M}_{\text{Lin. Reg.}}(m, L) = \underset{\{S \subseteq L \mid |S|=k\}}{\arg\max} \sum_{l \in S} \left(\hat{\beta_{l0}} + \sum_{i=1}^{|L|} \hat{\beta_{li}} \frac{|O_i^m|}{|O^m|} \right)$$

where, the expression in the parenthesis estimates probable occurrence distribution in locations for m.

Approach	Accuracy	Impact	Impact Diff.
Random	0.256	0.343	0.739
Greedy	0.296	0.372	0.76
Lin. Regression	0.328	0.241	0.626
Sharing Infl.	0.266	0.264	0.666
Transmitting Infl.	0.242	0.253	0.654
Spatial Infl.	0.373	0.309	0.685
Transmitting Infl. + Spatial Infl.	0.407	0.393	0.78
Sharing Infl. + Spatial Infl.	**0.421**	**0.403**	**0.789**

Table 1: Comparing the predictive models ($t_s = 5$ minutes, $k = 3$). The approach combining the community influence approach with spatial influence approach (*sharing influence + spatial influence*) performs the best.

5.2 Evaluation Metrics

We denote the best possible location subset that can be selected at t_s as $S^{m\star}_{t_s}$ ($S^m_{t_s}$ with a $\star$ on top). To evaluate the performance of the approaches proposed in this paper, we define three metrics:

Accuracy: This metric measures the similarity between the approximate subset, determined using our approaches, and the exact location subset that is determined after the completion of hashtag propagation. This measure is similar to other set comparison metrics like the Jaccard index. It is defined as:

$$Accuracy = \frac{S^{m\star}_{t_s} \cap S^m_{t_s}}{k}$$

where, k is cardinality of $S^m_{t_s}$. If the sets are identical, the accuracy is 1.0, and 0.0 if they are disjoint.

Impact: While accuracy measures the similarity between the sets, it doesn't measure the effect of selecting a particular subset over another. For example, it is possible that two disjoint sets of locations observe same number of occurrences after they are selected, resulting in the same impact. Hence, we also consider the subset **impact**, which measures the percentage of hashtag occurrences that were observed in the approximate location subset. It is defined as:

$$Impact = \frac{U^m_{S^m_{t_s}}}{|O^m \cup U^m|}$$

where, the numerator is the number of occurrences that were observed in $S^m_{t_s}$, after it was selected, and the denominator is the total number of occurrences of the hashtag. The impact value ranges from 0.0 to 1.0, with 0.0 signifying no impact, while 1.0 signifying maximum impact.

Impact Difference: If a hashtag is distributed uniformly across large number of locations, then the best impact for a given k might be small. In this case, the performance of an approach will be measured as low, even if it selects the best set. Hence, we can also measure the subset **impact difference** that measures the difference between the impact for the best subset and the approximate subset. It is defined as:

$$Impact\ Difference = 1 - \frac{U^m_{S^{m\star}_{t_s}} - U^m_{S^m_{t_s}}}{|O^m \cup U^m|}$$

Like the other two metrics, the lower the value of difference the better is the approach. A value of 1.0 signifies the impact is identical while a value of 0.0 indicates the subset has no impact at all.

5.3 Evaluating the Models

We now evaluate the performance of location subset selection approaches using the metrics defined in the previous section. We first evaluate the performance of the approaches for a fixed value of location selection time t_s and subset cardinality k. We then evaluate the performance of these approaches by varying the time used to select location subsets. Similarly, we then evaluate the performance of the approaches for different sizes of location subsets.

Experimental Setup: For our experiments we use two hashtag sets: (i) Training hashtag set, and (ii) Test hashtag set. The hashtag sets are extracted from Twitter hashtag propagations as described in Section 3.2. Techniques that require prior hashtag propagations (linear regression, sharing and transmitting influence) use the training hashtag set to build their models. For the spatial influence model, we set $\alpha = 1.01$.

We use the test hashtag set to evaluate the performance of the approaches. Given a hashtag from the test set, to evaluate an approach-metric pair, we replay the hashtag's propagation. At location subset selection time, we select location subset using this approach and then continue with the remaining propagation of the hashtag. At the end of this hashtag's propagation, we measure performance of the approach using this particular metric. We do this for all hashtags in the test set and calculate the mean score for this metric-approach pair. This experiment is done for a given value of t_s and k. We set $\beta = 0.5$ in (6) giving equal weight to both approaches.

Comparing the Models: We begin by fixing the selection time for each approach as 5 minutes (i.e., $t_s = 5$) and the number of locations to selects as 3 (i.e., $k = 3$). How well do the approaches predict future locations given only evidence of the first 5 minute's of a hashtag's lifetime? We report the results across all approaches for accuracy, impact, and impact difference in Table 1. Recall that accuracy measures the similarity between subsets selected by our approaches and the best subset; impact measures the actual percentage of occurrences observed in the locations; and impact difference measures the percentage difference between the best impact and the impact achieved using one of the approaches.

First, we observe that the approach combining the community influence approach with spatial influence approach (*sharing + spatial*) performs the best, with an accuracy of 42%, and impact of 40%, and an impact difference of 79%. Interestingly, we observe that approaches based on the spatial influence model tend to perform much better than approaches that use only historical hashtag propagations (e.g., linear regression). For example, the accuracy of the *spatial influence*, of *transmitting + spatial*, and of *sharing + spatial* is higher in all cases than all other approaches. We see similar strong results for the combined approaches (*transmitting + spatial*, and of *sharing + spatial*) as compared to all other approaches. Surprisingly, the community influence-based approaches alone (e.g., *sharing* and *transmitting*) perform the worst, even worse than the random and greedy approaches.

These results are significant because they illustrate the impor-

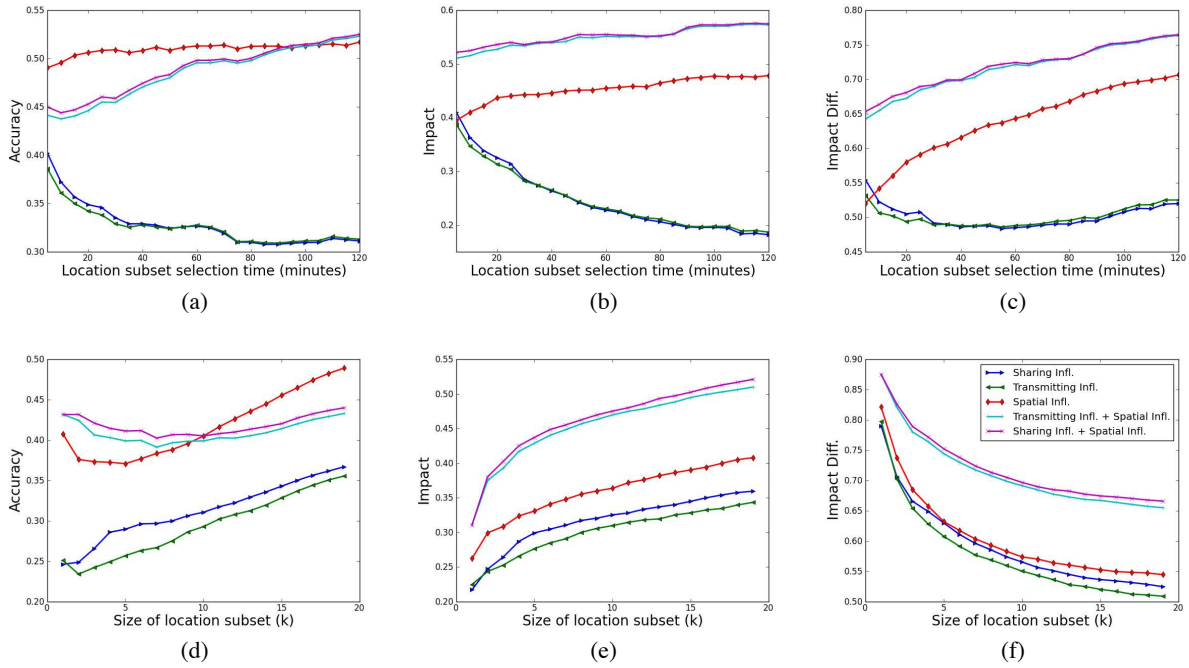


Figure 8: (Color) Varying the selection time (top) and varying the number of locations predicted (bottom).

tance of prioritizing the spatial influence model over the community affinity models, but also the combined power of incorporating community affinity into the spatial influence model for the best overall performance. Selecting future locations that will adopt a hashtags with very little knowledge of how a hashtag is going to propagate is a difficult problem. Based on these results, the performance achieved by the model that combines sharing probability with coverage probability is very encouraging. Most popular hashtags spread for several hours, but this model can identify 40% of all future occurrences of a hashtag within 5 minutes of the hashtag's first appearance. Also, the quality of locations selected by this model is high, as the locations it selected came close to 79% of the best performing locations.

Varying the Selection Time: What if we increase the time until the models have to make a prediction? That is, if we allow the hashtags to propagate for even longer, what impact does this have on the predictive ability of the models as they have access to additional evidence? Hence, we next varied the location subset selection time (t_s) from 5 minutes to 2 hours, keeping the k fixed at 20. We evaluated each approach for each selection time (e.g., after 5 minutes since a hashtag's first appearance, after 10 minutes, and so on up to 120 minutes) as shown in the top row of Figure 8. We plot the affect of varying the selection time against the five approaches, showing accuracy in Figure 8(a), impact in Figure 8(b), and impact difference in Figure 8(c).

We see that across all metrics, the approaches that use both sharing and transmitting influence coupled with spatial influence (the purple and light blue curves) improve with the increase in location selection time. As the time to select locations increases, each approach can observe a longer lifespan of a hashtag's propagation, leading to stronger evidence for making better predictions. In contrast, the community affinity approaches alone (*sharing* and *transmitting*, in blue and green) degrade in quality as the selection time increases (with a slight uptick for impact difference after 80 minutes). These results further confirm the importance of the spatial influence models as the single strongest predictor of hashtag spread.

An interesting result we observe in this figure is the performance of approach that uses spatial influence alone to select locations. We observe that the curve (red-diamonds) corresponding to this approach stays relatively constant irrespective of the value of t_s. This approach selects locations just based on spatial influence and hashtag distribution, hence a constant accuracy indicates that the probability scores for locations remain same irrespective of t_s, i.e., the overall probability distribution for a hashtag calculated after 5 minutes is similar to its probability distribution calculated after 2 hours. This result further strengthens our assessment, in Section 3.3, that early coverage for a hashtag is a good indicator of its final coverage.

Confirming the results from our previous experiment, we find that approaches that use the spatial influence model in concert with a community affinity model perform the best.

Varying the Number of Predictions: Finally, we evaluate the performance of each approach by varying the number of locations each predicts. Hence, we vary the cardinality k from 1 to 20, while fixing the selection time at 5 minutes, as shown in the bottom row of Figure 8. Across all three metrics – accuracy in Figure 8(d), impact in Figure 8(e), impact difference in Figure 8(f) – we again see the strong performance of the spatial influence models, both for the spatial model along (*spatial*) as well as the model incorporating community affinity into the spatial model (*transmitting + spatial* and *sharing + spatial*). As the number of locations increases, we see the accuracy of all approaches increase since each selects more top locations correctly. We also see an improvement in impact for all the approaches, with increasing cardinality. This result is straightforward since increasing the number of locations implies a higher number of occurrences are observed, which in turn increases the impact. But, the magnitude and rate for improvement of impact varies for all the approaches, with all the approaches that use spatial influence model showing greater impact than approaches that use community affinities only. This result is similar to the results observed in Figure 8(b). Finally, we observe that increasing the cardinality results in a decrease in impact difference for all approaches.

5.4 Summary of Results

Based on this experimental study, we find that:

- First, **distance does matter**. As shown in Table 1, we found that the spatial influence model – based on Tobler's first law of geography – is the single most important explanation of future hashtag adoption. Distance matters mostly because hashtags are fundamentally a *local phenomena*. Hashtags typically occur in an originating location and subsequently in nearby neighboring locations.
- Second, we additionally discovered that though the community affinity influence model alone performs worse than the spatial influence mode, **in combination** with the spatial influence model we can achieve the best fit for future hashtag adoption. This combination indicates that community affinities (like culture, language, and common interests) are a secondary factor

6. CONCLUSION

In this paper, we have begun an investigation of the global spread of social media. We have studied the geo-spatial properties of a collection of 755 million geo-tagged tweets and found that (i) pairs of locations tend to share common hashtags and adopt them around the same time, compared with locations that are distant; (ii) many distant location pairs, though, exhibit similar patterns of "closeness" in that they share hashtags and have a low hashtag adoption lag, suggesting some additional factor is "bending space"' to link the two locations; and (iii) the early occurrences of a hashtag are good indicators of the relative coverage of a hashtag's future spread (either compact or widely diffuse). Based on these observations, we developed two complementary models of hashtag spread – the spatial influence model and the community affinity influence models – and studied their effectiveness at predicting locations that will adopt hashtags in the future. We conclude that **distance does matter** as the single most important explanation of future hashtag adoption since hashtags are fundamentally local. We also find that community affinities (like culture, language, and common interests) enhance the quality of purely spatial models, indicating the necessity of adequately incorporating non-spatial features into models of global social media spread.

In our continuing work, we are interested in augmenting the developed models – that consider only the geo-spatial properties of hashtags – with additional evidence of the content of the hashtags (e.g., since politics-related social media may spread differently than sports-related social media) and with the underlying social network. Recall that the study in this paper has been completely orthogonal to the underlying social network and how social contagion affects hashtags spread. As part of this continuing work, we are interested in linking these geospatial diffusion models to these related efforts.

7. ACKNOWLEDGMENTS

This work was supported in part by NSF grant IIS-1149383, DARPA grant N66001-10-1-4044 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.

8. REFERENCES

[1] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. WWW '08.

[2] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. WWW '10.

[3] C. Bauckhage. Insights into internet memes. ICWSM' 11.

[4] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. WWW '12.

[5] F. Cairncross. *The death of distance: How the communications revolution is changing our lives*. Harvard Business School Press, Boston, 2001.

[6] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. CIKM '10.

[7] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Gonçalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. LSM '11.

[8] N. Dalvi, R. Kumar, and B. Pang. Object matching in tweets with spatial models. WSDM '12.

[9] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. COLING '10.

[10] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. VLDB '00.

[11] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10.

[12] 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.

[13] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. KDD '03.

[14] D. Kempe, J. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. In *IN ICALP*, 2005.

[15] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. KDD '08.

[16] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? WWW '10.

[17] K. Lerman and R. Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. *CoRR*, 2010.

[18] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. KDD '09.

[19] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. KDD '11.

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

[21] A. Oreskovic. Huffington post: Youtube video views hit 4 billion per day, Jan. 2012.

[22] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. WWW '11.

[23] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. ICWSM' 11.

[24] W. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(2), 1970.

[25] O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. WSDM '12.

[26] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. ICDM' 2010.