Spatio-Temporal and Events Based Analysis of Topic Popularity in Twitter

Sebastien Ardon[♦] sebastien.ardon@nicta.com.au
Amitabha Bagchi[*] bagchi@cse.iitd.ac.in
Anirban Mahanti[♦] anirban.mahanti@nicta.com.au
Amit Ruhela[*τ] aruhela@cse.iitd.ac.in
Aaditeshwar Seth[*] aseth@cse.iitd.ac.in
Rudra M. Tripathy[*] tripathy@cse.iitd.ac.in
Sipat Triukose[♦] sipat.triukose@nicta.com.au
IIT Delhi, India[*]
NICTA, Australia[♦]
C-DOT Delhi, India[τ]

ABSTRACT

We present the first comprehensive characterization of the diffusion of ideas on Twitter, studying more than 5.96 million topics that include both popular and less popular topics. On a data set containing approximately 10 million users and a comprehensive scraping of 196 million tweets, we perform a rigorous temporal and spatial analysis, investigating the time-evolving properties of the subgraphs formed by the users discussing each topic. We focus on two different notions of the spatial: the network topology formed by follower-following links on Twitter, and the geospatial location of the users. We investigate the effect of initiators on the popularity of topics and find that users with a high number of followers have a strong impact on topic popularity. We deduce that topics become popular when disjoint clusters of users discussing them begin to merge and form one giant component that grows to cover a significant fraction of the network. Our geospatial analysis shows that highly popular topics are those that cross regional boundaries aggressively.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data mining; H.3.5 [**Online Information Services**]: Web-based services

Keywords

Online Social Network; Topics; Diffusion; Events

1. INTRODUCTION

In the last decade, the microblogging service Twitter has attained a massive world-wide following with some estimates putting the user base at more than 1 Billion users. Nonetheless, fundamental questions remain unanswered. We know, for instance, that discussions around certain topics "go viral" whereas other topics die an early death. The network propagates some ideas, and some make no headway. In view of the enormous influence of online social networks (OSN), understanding the mechanics of these systems is critical. To characterize the properties of popular and non-popular topics is of surpassing importance to our understanding of how these complex networks are shaping our world.

In this paper we present a large-scale measurement study that attempts to describe and explain the processes that animate microblogging services. We study a large set of popular and non-popular topics derived from a comprehensive data set of tweets and user information taken from Twitter. A key strength of our study is that we observe both popular and not-so-popular topics. This allows us to hypothesize about the temporal and spatial behavior of popular topics and support our hypotheses by showing that non-popular topics display contrary behavior.

The following simple abstraction inspired by the study of infection spread in networks underlies our work: We view each topic as a kind of organism that replicates when a user tweets about it. We study the spread of the topic through the network, tracking the structure of the infection's growth, in the way that epidemiologists do for infections. Drawing on ideas from the study of epidemiology, complex systems and statistical mechanics we try to characterize popular and unpopular topics in terms of their patterns of growth and decay and provide some hypotheses on what differentiates a popular topic from an unpopular one. The major difference between our work and the models of virus spread is that here an individual tweet may have more than one topic, e.g. a tweet on Angelina Jolie undergoing a mastectomy may have topics "Angelina Jolie" and "Breast Cancer" and may lie in the intersection of the spread of these two topics.

Our work emphasizes the structural aspects of topic spread. We give the semantic aspect its due importance in the process of topic identification, using a laborious manually driven methodology to ensure that our topics make sense, and then proceed to study the fundamental temporal and spatial aspects of the spread of topics which is our main focus. In particular, we study topic movement over two interrelated spatial dimensions: the topology of the Twitter network as formed by "follower" and "following" relationships, and the geospatial embedding of that network in the map of the world.

Our study spans several aspects of spatial diffusion on Twitter, but our primary focus is on characterizing the temporal and spatial underpinnings of popularity. We focus on three important aspects as described in the sequel. First, in

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505525.

Section 4, we study the effect of topology and the dynamics of topic spread on popularity. The primary objects of study to this end are the subnetworks formed by users discussing each topic. While it is known that the Twitter network, like most large OSN, contains a giant connected component, a key finding is that the subgraph of users talking about a popular topic on a particular day always contains a giant connected component containing most of the nodes (users) of the subgraph, whereas the subgraphs of non-popular topics tend to be highly disconnected. To summarize, we make the following observations:

Hypothesis 1. Most of the people talking about a popular topic on a given day tend to form a large connected subgraph (giant component) while unpopular topics are discussed in disconnected clusters.

Hypothesis 1a. The giant component forms when many tightly clustered sets of users discussing the topic merge.

Second, we study the impact of geography on popularity by partitioning the Twitter network according to regional divisions and studying the behavior of popular and non-popular topics.

Hypothesis 2. Popular topics cross regional boundaries while unpopular topics stay within them.

Finally, we study how topic initiators influence popularity of the topic, and make the following observations:

Hypothesis 3. Twitter is a partially democratic medium in the sense that popular topics are generally nourished by users that have large numbers of followers; however, for a topic to become popular it must be taken up by users having scant followers count. Further, regions with large number of heavily followed users dominate Twitter.

Since topics last for long times and often witness surges in activity due to some occurrence in the real world, we identify events within each topic using a methodology based on the inter-arrival time of tweets. We use the term *event* for these surges and we divide them into five distinct phases. This allows us to validate our hypotheses in an aggregated manner by studying our claims within a large set of events rather than topics which could become popular then fall in popularity, then rise again within the time window covered by our data set, thereby skewing our results.

Apart from the highlights mentioned above, we review related work in Section 2. We describe the various methodological issues that needed to be surmounted to perform our study in Section 3. Section 8 concludes the paper with a discussion of the implications of our observations on different aspect of the OSN sphere.

2. BACKGROUND AND RELATED WORK

Leskovec, Backstrom and Kleinberg's seminal work on the evolution of topics in the news sphere was the starting point for this paper [11]. They studied how the growth of one topic affects the growth of other topics in the blogosphere. They identified and tracked a small number of popular threads, and showed that the growth of the number of posts on a thread negatively impacts the growth of other threads. The basic question that arose on reading that work was this: Can the nuances of the temporal evolution of topics be explained by a more thorough study of their spatial evolution? Working with a data set taken from Twitter we were able to extract the high level of structural and geographical information about the actors of the process that has allowed us to answer this question in the affirmative. This allows us to challenge the line of research that studies only the temporal evolution of topics [21], or seeks to explain this evolution on the basis of content [20].

Following the paper cited above there has been more interest in understanding how information and ideas propagate on OSNs. A pioneering study on these phenomena on Twitter was conducted by Kwak et. al. [9] where several aspects of topic diffusion were studied. Of particular relevance to our work was their study on the topological properties of retweet trees. Since our data set is built on the data set they used (cf. Section 3) for details), our work can easily be compared. Our major contribution is that we work with a more general notion of a topic and that we work with an ecosystem of topics. Also our work views the diffusion of topics through the lens of what we call "topic graphs" (cf. Section 4), that are a significant generalization of retweet trees. Retweet cascades have also been studied specifically for the case of tweets with URLs in them by Galuba et. al. [5] and by Rodrigues et. al. [14]. There is a line of work that seeks to uncover the structural processes behind topic diffusion by studying cascade models (e.g. Ghosh and Lerman [6], Sadikov et. al. [16]) but we feel this is a limited view of the effect of topology and try to view the network structure in a more complex way. Myers et al. [13] measure the importance of external effect in information diffusion in social networks. By modelling an information diffusion models with external effects and also doing a large scale measurement analysis on Twitter they have showed that, around 70% of the information volume is due to network effect, while the remaining 30% is governed by the external effects.

In another work, Sousa el al. [17] investigated whether user interactions on Twitter are based on social ties or on topics, by tracking replies and message exchange on Twitter; their study is focused on only three topics namely sports, religion, and politics. Romero et al. [15] studied topic diffusion mechanisms on Twitter by focusing on topics identifiable by hashtags. They study the probability of a topic adoption based on repeated exposures, and provide quantitative evidence of a contagion phenomenon made more complex than normal studies of virus-like phenomena by the existence of multiple topics, and briefly report on the graph structure of topic networks. More recently, Lehmann et al. [10] study the dynamics of hashtags in Twitter. They found exogenous factors more important than endogenous factors to make a hashtag popular. In another recent work, Weng et al. [19] analyze how competitive memes diffuse in social network. They find that the combination of social network structure and competition for finite user attention is a sufficient condition for the emergence of broad diversity in meme popularity, lifetime, and user activity. They have developed an agent-based toy model of meme diffusion and compare its predictions with the empirical data. For the measurement work they use Twitter data where they consider hashtag used in the tweets as memes. One major limitation of these works we found is that only a very small fraction (approx. 10%) of tweets are tagged with hashtags. Our methodology

of using a Natural Language Processor (OpenCalais) allows us to study topic diffusion on a much larger scale than in this work since our topic choices are not limited to hashtags. In a similar work Asur et al. [1] analyze the trending topics of the Twitter network. Their work mainly focused on the temporal evolution of trending topics. Compared to our work, we have analyzed both popular topics as well as non-popular topics and also our topics section model is different. In addition to that, we have showed temporal, spatial and geographical evolution of both popular and non-popular topics.

On the geographical front, Yardi et al. [22] examine information spread along the social network and across geographic regions by analyzing tweets related to two specific events happening at two different geographic locations. As an aside we mention that Krishnamurthy et. al. characterized the geographical properties of the Twitter user base in 2008 [8].

On a more general level, we note that it is implicitly assumed that the attention of users on a platform like Twitter is elastic but bounded (see e.g. [12]) and hence, the diffusion process is essentially a competitive one, even if it is not explicitly adversarial. The study of competitive diffusion has largely revolved around the application domain of viral marketing where there is competition between different products [2, 4, 7]. Budak et al. [3] consider the problem of diffusion of mis-information, where opposing ideas are competing and propagating in a social network. The study of processes by which rumor spread may be combated is another example of competitive diffusion [18]. Our work provides an important input into this area of study, articulating the properties of a complex system that requires extensive study to model correctly and comprehensively.

3. METHODOLOGY

3.1 Data Set Description

Our primary data set is a snapshot of Twitter activity over a period of 3 months from June to August 2009. This snapshot comprises 196 million tweets from 10 million users. We were able to tag a subset of these tweets with topics via a manually overseen process that we will explain. Within the tweets tagged by a single topic we identified "events" using the inter-arrival time between tweets as a metric to identify the different phases of an event. We also reverse-mapped the topic-tagged tweets to a network of follower-following relationships between the users. The location of these users was identified where possible. We now discuss each aspect in greater detail.

The tweets.

Our tweets came from a portion of the 'tweet7' data set crawled by Yang et al. [21] between June and August 2009.

The network.

Follower-following relationship between users were taken from a data set collected from Twitter during the same time period by Kwak et. al. [9]. We filled in the gaps by crawling the profile info of the users who had tweeted in the tweet7 data set but were not present in [9]. Both the in-degree and out-degree distributions follow a power-law as mentioned in [9]. Most users follow a few people and a few users

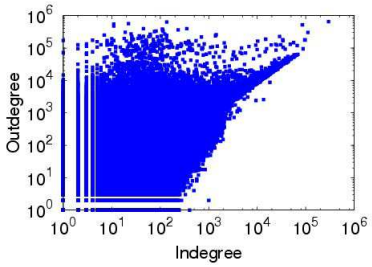


Figure 1: In-degree vs Out-degree in the Twitter network

are followed by a large number of people. We notice that the in-degree and out-degree values are positively correlated, as shown in Figure 1 by an overall clustering of points around the $x = y$ line. In addition, we observe an additional cluster of points on the top-left quadrant of this graph, which are users with very large numbers of followers.

Locations.

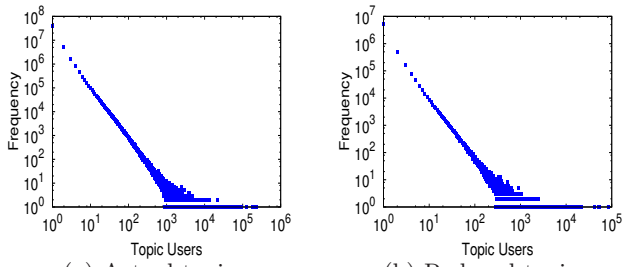
Using the Twitter User ID available in the tweet7 data set, we queried Twitter using its API to extract location information from user profiles. A number of users had their location specified in latitude and longitude (which is often done by certain GPS-equipped mobile devices while using Twitter). We passed this information to the Yahoo! Placefinder service which resolved it in terms of City, State and Country. For other users we passed the text in their profile's "Location" field to Yahoo! Placefinder. In many cases the service was either unable to resolve the location and in some other cases it provided a list of possible matches with associated scores. In the latter case we took the location with the highest score as the user's location. The process was successfully completed for about 60% of the users.

The largest number of users in the data set are from USA (57.6% users), followed by UK (7.7% users), Brazil (7.1% users), and Canada (3.7% users). For our geographical topic diffusion analysis (Section 5) we further sub-divided USA into five commonly acknowledged regions: Northeast (10.7%), Southeast (13.5%), Midwest (10.4%), Southwest (6.8%), and West (16%) where each sub-region differs with other sub-regions in terms of history, traditions, economy, climate, and geography.

Topics.

Using hashtags to identify topics in tweets has been the norm in the literature (e.g. [15, 9]), but sparsity is a problem. In our dataset only 10% of the tweets contain a hashtag. We took these hashtags as topics but also augmented the topic set by a laborious combination of automated and manual intervention. We began by bunching tweets into files of size 40KB and passing it to the OpenCalais text analysis engine to extract entities, topics, places and other such tags. On receiving the output we then went back over each tweet in the bunch and associated OpenCalais's output tags with them by simply string matching the tag's content with the tweet. Each tweet was allowed to have multiple tags. We used URLs as topics but we *did not* follow the URLs to the respective webpages obtain further topics since that would be prohibitively expensive. We obtained 48 million topics for 114 million tweets which includes the topics returned by OpenCalais, Hashtags and URLs. The remaining twee-

ts were discarded. The popularity distribution of these 48 million topics follows a power-law shape (Figure 2(a)): most topics are talked by very few users (< 10 users). To make a manageable set of topics, we exponentially scale down the set of topics to 5.96M taking care to include both popular and non-popular topics. Figure 2(b) shows the frequency distribution of this reduced topic set. We can see in Figure 2(b) that the frequency distribution follows a power-law. To identify and measure the popularity diversity of topics,



(a) Actual topics (b) Reduced topics

Figure 2: Frequency distribution of topics.

Figure 3 compare the number of users to the number of tweets, by plotting those two variables on a scatter plot, for each of the 5.96M topics. This graph effectively shows the difference in popularity for all topics. From the graph, we can see that users of unpopular topics typically tweets more than one time on that topic. Popular topics on the other hand, are typically tweeted once by most users. Finally, we

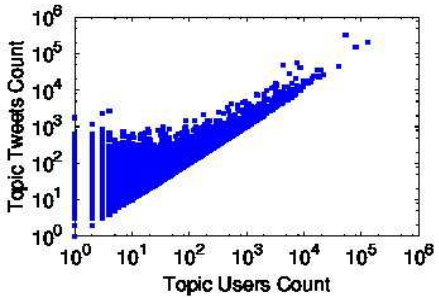


Figure 3: Topic popularity diversity

removed all topics that had been tweeted on by less than 15 users, and then manually examined a sample of 6000 topics drawn randomly from the 5.96M topics in such a way that the power-law distribution was maintained (i.e. we sorted the topics by rank, made uniform sized buckets and picked a fixed number from each bucket). After merging duplicates and removing nonsensical topics we were left with 4135 topics. A classification of these topics that we will use separates them into 3 categories. *Popular topics*: tweeted on by at least 10,000 users; *Medium popular topics*: tweeted on by at least 1,000 and at most 10,000 users; *Non popular topics*: tweeted on by at most 1000 users.

Events.

A topic, typically an identifier for a person, place, object, brand, occurrence etc. (e.g. IRANELECTION, BARACK OBAMA, INDIANA, IPHONE), has a lifetime during which several minor and major happenings cause surges in tweeting activity about that topic. We use the term "event" for such a surge, and focus on the inter-arrival time (IAT) of tweets to identify these events. Our model on event has a lifetime that can be partitioned into five divisions according to Figure 4. The first phase of the event is the *pre-event Phase* (t_1 to t_2) in which initiators introduce the topic into the network. The second phase of the event is the *growth phase* of an event (t_2 to t_3) in which early adopters talk on the topic. The third phase of an event (t_3 to t_4) is the *peak phase* in which an "early majority" of people talk on the topic. The fourth phase of an event (t_4 to t_5) is the *decaying phase* in which a "late majority" of people discuss the topic. The last phase of the event (t_5 to t_6) is the *post-event* phase in which "laggards" talk on the topic. To find events within

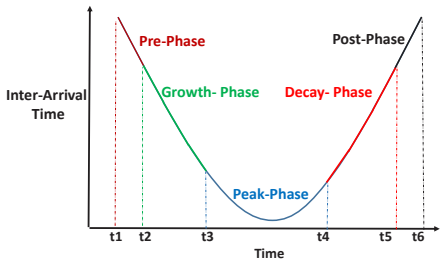


Figure 4: Phases of an Event

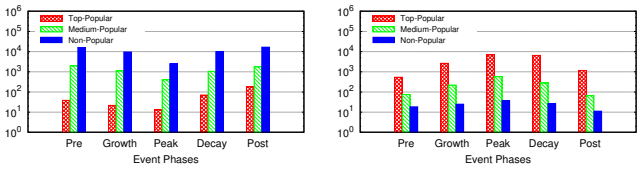
a topic we compute the inter-arrival times (IATs) between tweets. The sequence of IATs is smoothed using a median filter. As indicated by Figure 4, the IAT peaks at the end of an event so we detect the peaks in the smoothed sequence and demarcate the region between subsequent peaks as events. A fundamental problem with this method is that it is sensitive to noise and so it did not yield good results for topics with less than 100 tweets. As a result we considered only topics with more than 100 tweets. This constraint reduced our set of 4135 manually inspected topics to a very small number so we had to go back to the 5.96 million uninspected topics and choose those topics that had at least 100 tweets in them. These were approximately 8000 in number. A manual inspection of 1000 of these topics showed that these were topics that largely made sense.

We classified events into three categories. *Popular events*: contained at least 10,000 tweets; *Medium popular events*: between 500 and 10,000 tweets; *Non-popular events*: between 100 and 500 tweets. Out of 16492 events detected 93.29% events were non-popular, 6.54% events were medium popular events and 0.15% events were popular events. As expected the popular events largely belongs to popular topics whereas a large proportion of non-popular events belongs to non-popular topics.

Figure 5(a) shows that the mean inter-arrival time for popular events is two orders lower than that of non-popular events. For all three events categories, the mean IAT is highest when the topic starts, decreases when the topic grows, and is least when the topic is at the peak. The mean IAT increase again when the topic goes in event-decay phase and increase further when the topic goes in post-event phase.

3.2 Three graph structures

A common, and very difficult, question in the study of diffusion in social network settings is the *attribution* question i.e. when a user performs an action, what has caused him or



(a) Mean IAT (b) Mean # Users

Figure 5: Event phases: IAT and User counts

her to do so. In a social setting we work with the assumption that a significant fraction of the actions performed by a user can be attributed to actions provided by those other users whose actions the user observes: in the case of Twitter this is the set of users who he or she follows. In the topic diffusion setting it is perhaps impossible to construct the true graph of which tweet was influenced by which other tweet (although in virus spread it may be possible to determine using DNA based techniques which organism reproduced to give rise to a particular organism). Hence, we try to approximate the process of attribution through two graph structures that definitely contain this process: First, we study the *lifetime graph* of a topic; this is the subgraph induced on the Twitter network by all the users who have tweeted on that topic at any time in our window. Secondly, we study the *cumulative evolving graphs* of a topic. We denote by $G_i^t = (V_i^t, E_i^t)$, the cumulative evolving graph for topic t on day i and define it as follows:

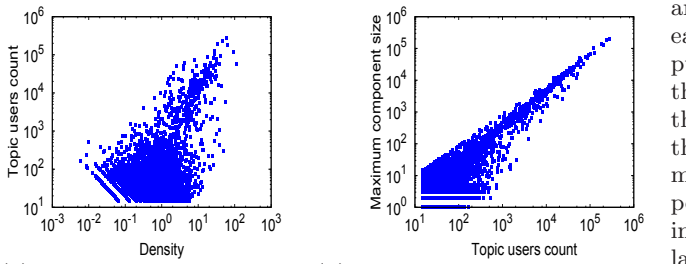
- The vertex set of G_0^t comprises the users V_0 who tweet about t on day 0. The edge set is empty.
- The vertex set v_i^t of G_i^t is the set of all users who have tweeted on a topic in days 0 through i. An edge $(u \rightarrow v)$ is added to E_i^t if $u \in V_{i-1}^t$ and v tweets about t on day i.

Clearly if user A follows user B and is influenced by a tweet from user B and tweets on the same topic, this relationship will be contained in both the lifetime graph and the cumulative evolving graphs.

We also study a snapshot graph, we call it the *evolving graph* of a topic. In particular, we partition the tweets related to a particular topic by day and for each day we construct the subgraph induced on Twitter by the users who have tweeted on that topic on that particular day.

4. PROPERTIES OF TOPIC GRAPHS

In this section we establish Hypothesis 1 and argue towards Hypothesis 1a.



(a) Density vs popularity (b) Component Size vs popularity

Figure 6: Lifetime graphs

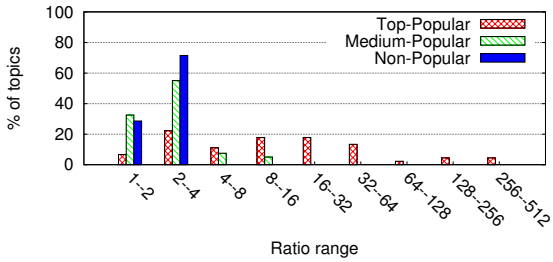
4.1 Lifetime Graphs

We constructed lifetime graphs for each of the topic of our reduced dataset. Our first observation from these graphs is that popular topics tend to occupy the more well-connected portions of the network. To establish this we studied the relationship of the total number of users who have tweeted on a topic (referred to as the *topic user count*) to the *density* of the lifetime graph of the topic (defined as the number of edges per user in the graph). In Figure 6a, we note something important: the lifetime graphs of non-popular topics (e.g., user count < 1000) do not have densities greater than 10, and in fact many tend to have a density less than 1. A density of less than one for a subgraph of a reasonably well-connected graph like the Twitter network clearly indicates a high number of small isolated clusters. This isolation is observed even in the lifetime graph which establishes relationships between users even where they may not exist, for example, by putting an edge from u to v although v may have tweeted on the topic *before* u. Hence, Figure 6a strongly supports one side of Hypothesis 1: less popular topics generally exist in highly disconnected clusters.

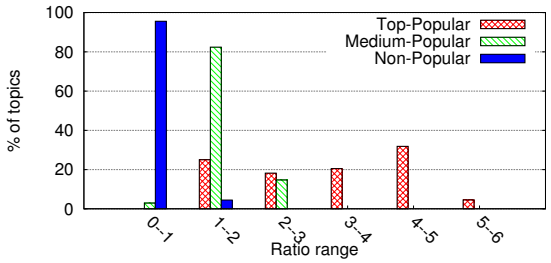
It is difficult to establish the other direction of Hypothesis 1 from lifetime graphs because of the optimistic selection of edges mentioned earlier. Nonetheless we get strong indicative evidence for our Hypothesis that a popular topic tends to be discussed in one large cluster that contains most of the users that have tweeted on that topic. This evidence comes from studying the relationship of the topic user count to the size of the largest connected component of the lifetime graph, as shown in Figure 6b. From Figure 6b, notice that for more popular topics there is a clear linear relationship between the popularity of the topic and the size of the largest component of the lifetime graph. This is strongly indicative of Hypothesis 1, although it cannot be used as conclusive evidence.

4.2 Evolving Graphs

It is in the study of evolving graphs that we are able to establish that most users tweeting on popular topics form one large connected component (we will refer to this large component as the *giant component* from now on). For each topic category–popular, medium popular and non-popular–we randomly chose 40 topics, and computed evolving graphs for each. For each day's graph, we then computed the ratio between the sizes of the largest and the second largest component, and also the ratio between the radii of the largest and second largest component. In Figure 7(a) and (b), we present histograms for the ratios of component and radii sizes. The buckets divide the ranges of ratios observed for the size and the radius. We find the median ratio for each topic and display the percentages of these medians that land in each bucket. Note that only the highly popular topics populate the buckets with size ratio greater than sixteen, and that the median size ratio for these popular topics goes all the way up into the range of 10^2 and this is just the median, the maximum tends to be much higher but we study the median here because it is a more robust statistic. Most unpopular topics stay below 4 showing a remarkable evenness in the distribution of component sizes. The radii ratios similarly show that the width of the reach of the popular topics comes from the width of one large component rather than from a large number of small components. This effectively establishes Hypothesis 1.



(a) Ratio of components size



(b) Ratio of components radius

Figure 7: Median value of the ratios

Moving towards Hypothesis 1a, we first clarify in the context of evolving graphs what we mean when we say clusters merge. If we visualize the social network as a set of communities connected through users who may belong to multiple communities, our narrative of topic spread says that topics that are going to become very popular witness intense discussion *within* communities at first. When the level of intensity rises then the users who bridge communities enter the discussion in a big way causing a merging of what were earlier disjoint discussions. If Hypothesis 1a is correct then it can be reinterpreted to mean that the bridge users serve as a barometer of the topics rising popularity. To investigate the applicability of this narrative we study the conductance of evolving topic graphs. We define the *conductance* $\phi(S)$ of a subset of nodes S of a directed graph $G = (V, E)$ as the ratio of the edges outgoing from the vertices of S that land outside S:

$$\phi(S) = \frac{|\{(u \to v) : u \in S, v \in V \setminus S\}|}{|\{(u \to v) : u \in S\}|}.$$

Clearly, the higher the value of $\phi(S)$, the more the number of nodes outside S that are made aware of a topic being tweeted by the users in the set S. In Figure 8, we plot the evolving value of the conductance of the user set of the day's graph alongside the evolving topic user count for four topics: one less popular topic "CAMBRIDGE", one periodically popular topic "FOLLOWFRIDAY", and two topics that display distinct and very high peaks in their popularity "MICHAEL JACKSON" and "IRANELECTION". Observing the three popular topics we notice that conductance is very high just before the peak is seen. As soon as the peak is formed the conductance dips down to a low value. This supports Hypothesis 1a because when the users that bridge distinct clusters start tweeting on the topic then a larger number of edges become internal to the day's topic graph, hence the conductance should dip as it does. Again, we clarify that this result is merely indicative of Hypothesis 1a.

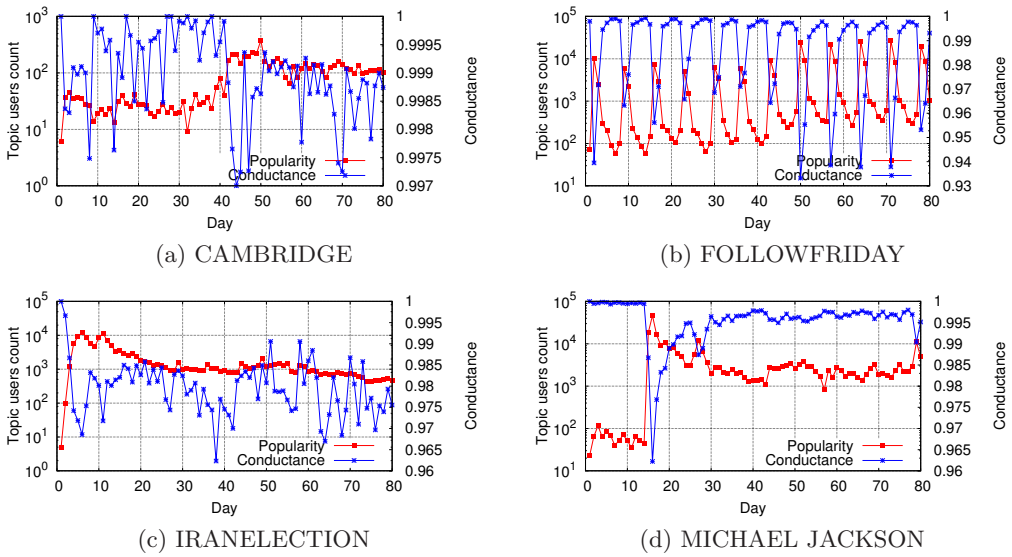
There are a number of other interesting artifacts that can be observed here. The sharp peak in Figure 8(d) comes on the day of Michael Jackson's demise. The conductance for this topic was uniformly high earlier, indicating a steady level of discussion about Michael Jackson, in tune with his general popularity. But his death leads to a sharp rise in tweets about him, causing an immediate dip in the conductance. After this initial dip the conductance rises again but no peak comparable to the first one is see, indicating that a high sustained level of interest in this topic is accompanied by a high sustained level of disinterest in the followers of the users continuing to tweet about Michael Jackson. Figure 8(c) shows a similar initial behavior accompanying an event, the holding of elections in Iran. Subsequently there is sustained discussion which is more of the nature of a conversation (the latter part of the "IRANELECTION" trajectory shows an unusually high number of tweets per user on this topic). This conversation proceeds in regions of the network that have reasonably high conductance but occasionally show dips in conductance, indicating a higher level of clustering in the user set, something that might be expected of a conversation. We note the the similarly high values of conductance displayed by the topic "CAMBRIDGE" in Figure 8(a) have a different connotation to the high values seen in the other graphs because, like most less popular topics, this too shows highly disconnected daily graphs.

4.3 Cumulative Evolving Graphs

By using a timing relationship to establish edges for the construction of cumulative evolving graphs, we make them a better approximation for the spread of a topic than the lifetime graphs we studied in Section 4.1. In Figure 9, we plot the fraction of nodes in the largest component of the cumulative evolving graph for two highly popular topics "MICHAEL JACKSON" and "IRAN ELECTION" and two less popular topics "SMARTPHONE" and "INDIANA, UNITED STATES". Note that even at the end of their evolution the two less popular topics have only 25% and 35% of their users in the giant component of the cumulative evolving graph while the two popular topics have half the users in the giant component even before the time window finishes. This supports Hypothesis 1 since the cumulative evolving graph is a better approximation of the spread of a topic. But, more importantly, the sharp rise in the fraction of nodes in the giant component that accompanies a peak in the evolution of the number of tweets stands in support of Hypothesis 1a because a merging of smaller clusters into one large cluster would be accompanied by a sharp rise in this fraction. It could be argued that this rise in the fraction is because of a sharp growth in the number of users in the largest component rather than a merging of clusters, but that seems unlikely given the extent of the rise, and the large number of users already present in the cumulative evolving graph at that point. The less popular topics also show increases in the fraction when their topic user counts drift upwards, but the rise is much less dramatic than that shown by the popular topics, and could possibly be explained by a general growth in the larger component rather than a radical merging of smaller clusters.

5. GEOGRAPHICAL ANALYSIS

This section establishes Hypothesis 2. We argue that the popularity of topics is correlated with their geographical spread. We begin by simply studying the number of re-



(a) CAMBRIDGE (b) FOLLOWFRIDAY

(c) IRANELECTION (d) MICHAEL JACKSON

Figure 8: Evolving graph conductance.

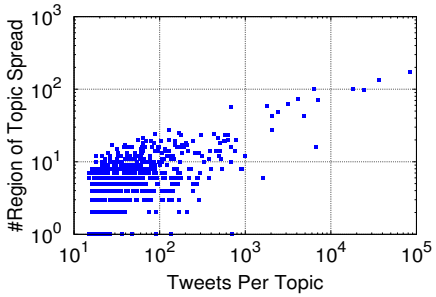


Figure 10: Users vs regions count.

gions represented by at least one user talking about a topic and plotted it against the popularity of the topic (see Figure 10). It is quite clear from this plot that the number of regions touched by less popular topics is less than those touched by more popular topics. This plot does not establish our Hypothesis but it is indicative of it in the sense that it does not falsify it either. In order to establish the Hypothesis, we investigated a geographical property of the cumulative evolving graphs defined in Section 4. For each topic we determined the fraction of edges in the cumulative evolving graph that went from one region to another; that is, we studied the fraction of edges $(u \rightarrow v)$ such that u belongs to one region and v is a user from another region. The evolution of this fraction for three topics, one popular, one medium popular and one non-popular (as defined in Section 3.1) is shown in Figure 11. We observe that the highly popular topic "BARACK" shows a high fraction of edges crossing regional boundaries throughout its evolution, ranging between 0.74 and 0.81. On the other hand the topic with medium popularity, "CAMBRIDGE", has a low fraction of edges crossing regions. It's noteworthy that an increase in the popularity of the topic "CAMBRIDGE" is accompanied by an increase in the fraction of edges crossing regional boundaries. This further supports Hypothesis 2. The topic "HAMBURG" which has low popularity shows a very small fraction of edges crossing regional boundaries.

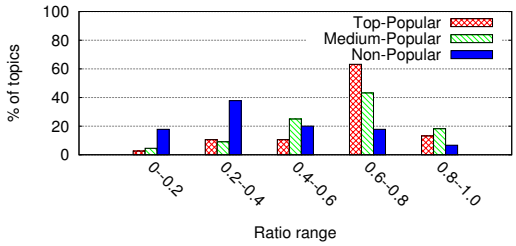


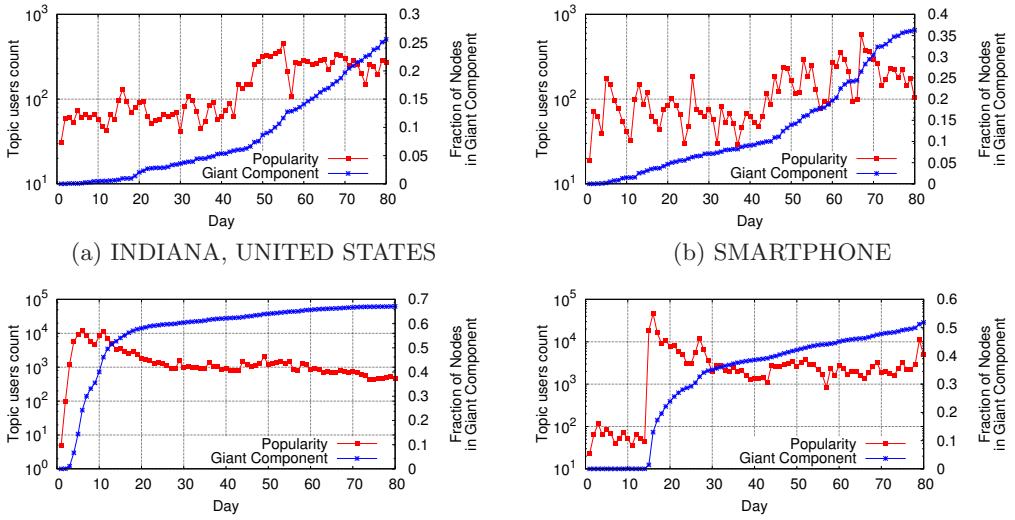
Figure 12: Fraction of edges cross regional boundaries

We also took 40 topics from each category (as we had done in Section 4) and computed the mean and median of the fraction of edges crossing regional boundaries for the entire period in our window where the topic is tweeted on. We plotted a histogram using five different ranges for this fraction (see Figure 12). This histogram clearly shows that the most popular topics tend to have a very large fraction of edges crossing regional boundaries while the least popular topics have cumulative graphs that generally evolve within regional boundaries with small fractions of edges going to other regions.

6. THE INFLUENCE OF INITIATORS

The sudden rise in importance of Twitter as a global communication medium has made it important to study who are the entities that wield most influence on this medium. In this section we make a finding, as stated in Hypothesis 3, that popular topics are generally nourished by users that have very high follower counts. These users are usually either news media outlets or media personalities (pop stars, politicians, writers etc.). Finding that the mean number of followers of a user in our data set's Twitter network was 65.7 and the standard deviation of this quantity was 1291.7, we decided to designate any user with more than 3,000 followers as heavily followed user.

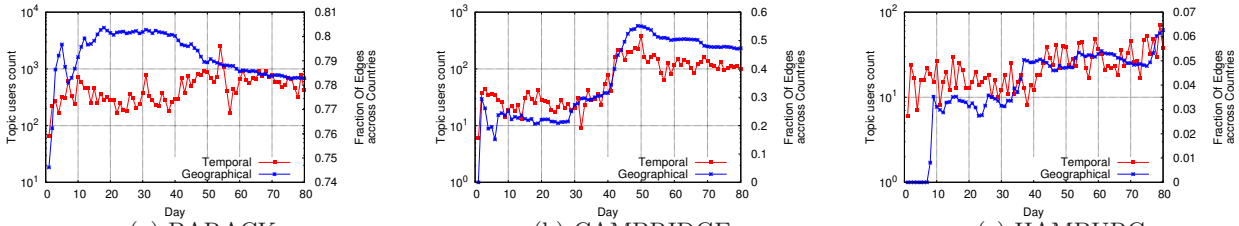
We study the influence of initiators on the number of users tweeting about a topic on a day with the filtered set of topics (i.e., those whose first tweet appeared at least 7 days after



(a) INDIANA, UNITED STATES (b) SMARTPHONE

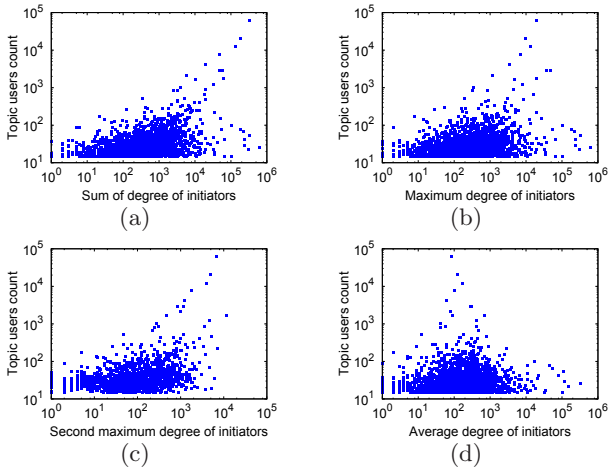
(c) IRANELECTION (d) MICHAEL JACKSON

Figure 9: Giant component evolution over time.



(a) BARACK (b) CAMBRIDGE (c) HAMBURG

Figure 11: Popularity vs edges crossing geography.



(a) (b) (c) (d)

Figure 13: #followers of initiators versus Users

the beginning of our data set's time window). In the Twitter network, there is no rule by which we can find the initiators of the topics since, apart from explicit retweets, we do not know when a user is influenced by a tweet. Hence, we use the following heuristic: we consider the early users (initial 5% of all tweets on a topic) as the initiators of the topic.

Figure 13 (a) shows the relationship between popularity and the total number of followers of the initiators. Note that highly popular topics have very high aggregate followers of the initiators. Figures 13 (b) and (c) show that celebrity users were indeed involved in initiating highly popular topics while most unpopular topics were initiated by users with a low number of followers. However, Figure 13(d) shows that the average number of followers of the initiators of highly popular topics is in the hundreds rather than the thousands, indicating that there are some initiators with relatively small number of followers involved in these popular topics.

An interesting observation can be made by looking at the points plotted near the bottom right corner of plots in Figures 13. These are topics started by a few celebrities that did not achieve any popularity. Hence, we see that while it is the case that celebrities drive the popularity of topics, it is not the case that every topic promoted by celebrities becomes popular. This helps us establish Hypothesis 3: Celebrities influence the spread of topics, but cannot make a topic popular unless common users pick up on them.

We expect that regions containing larger numbers of Twitter users will influence the topics discussed to a greater extent. To determine this we tabulated the number of topics for which each region has at least one user in the initiator set (cf. Figure 14a). We observe that an increase in users in a given region could potentially lead to an increase in the share of the topics initiated therein. However, a cautionary note is struck in Figure 14b which says that the countries which have large number of heavily followed users initiate large number of topics. Further, we did not consider all

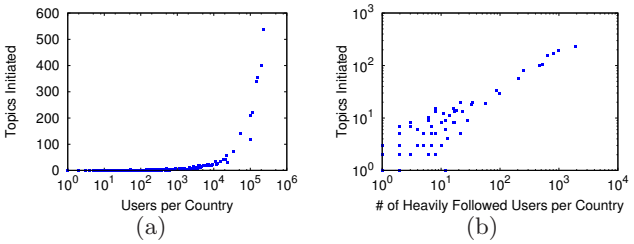


Figure 14: Geographical analysis of topics initiated.

topics for this plot, instead focusing on only the top 500 topics (by topic user count) in the filtered set. We see a kind of continuity with the Hypothesis 3 here. The cultural and political dominance of certain regions that existed before Twitter came into being is reflected in the presence of a greater number of celebrity users in those regions, and consequently translates into a greater impact for those regions in terms of popular topics.

7. AGGREGATE ANALYSIS USING EVENTS

In section 4 to 6, we have validated the three Hypothesis using sample of topics. In this section, present an aggregated analysis performed on events–surges in twitter chatter about a particular topic–to demonstrate that our hypothesis hold over a larger set and not just on the sample we presented earlier. As explained in Section 3.1 we detected 16492 events from 8250 topics. We have studied the evolving graphs of these events.

We first buttress the results of Section 4.2 by plotting a histogram of the ratio of size of largest to the second largest component of each event (Figure 15). The solid bar denotes the mean of the ratio and the narrow line in the middle of the bar shows the extent of one standard deviation around the mean. We note that popular topics see much larger ratios of the largest to the second largest component (with a fairly small standard deviation in the peak phase). For non-popular topics this ratio does not cross 8 even during the peak phase. This again validates Hypothesis 1. In Figure 16,

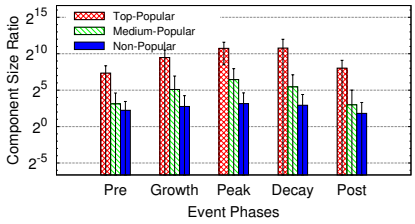


Figure 15: Largest to 2nd largest component size ratio

we have shown the mean value of conductance of the three events classes at various event phases. We observe that for popular events the conductance dips sharply as we move towards the peak phase, which is indicative of the merging we talk about in Hypothesis 1a, and then rises again as the topic decays, indicating that the graph is again falling into disconnected components. To validate Hypothesis 2, we determined the number of regions in which the topic associated with each event has been discussed. In Figure 17

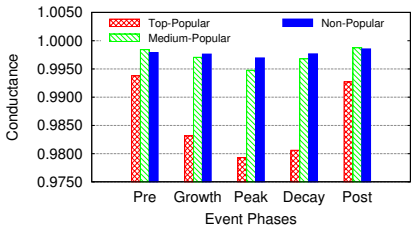


Figure 16: Conductance

we see that there is a significant growth in the number of regions where the topic is discussed as a topic moves into the peak phase of a popular event, whereas the mean number of regions remains more or less flat for non-popular events. The standard deviations of all these values (indicated in thin lines in the middle of each bar) indicate that the means are a good estimate of the aggregate behavior.

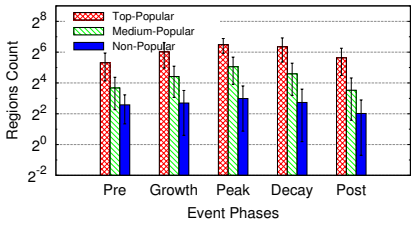


Figure 17: Number of regions

8. CONCLUDING REMARKS

The studies we have presented in this paper have wide-ranging implications, some of which, we hope, will be discovered in the future. For now we present a brief discussion of those area we feel our results may have an impact on.

Perhaps the most important implication pertains to the role and impact of highly influential users (and consequently of highly influential geographies). The rise of OSNs has been accompanied by a triumphal narrative of democratization of communication through technology, and while it is true that Twitter and other OSN platforms have played an important role in giving voice to individuals who might otherwise find it difficult to speak to an audience beyond their immediate geography, our study shows that traditional holders of power and influence have not been unseated.

Our hypothesis on how a giant component forms on Twitter–by the merging of smaller tightly clustered sets of users–is an important input into the sociology of how information is transacted on a social network. There is reason to believe that despite the fact that OSN platforms bring the world closer, older notions of proximity and community continue to contribute significantly to popularity in the way described. Our study is broad in nature and captures a coarse phenomenon that we hope will excite sociologist and invite them to tease out the finer nuances that lie within such phenomena.

From an engineering standpoint issues of content distribution and caching can be addressed from observing that highly popular topics cross national boundaries. A closer study of which national boundaries are crossed more often than others could underpin efficient content placement methods.

Our results could also be of great interest to those involved in using the vast reach of media like Twitter to advertise their products and services. The notions of trust and reputation inherent in OSNs have been leveraged to a great extent already for marketing purposes. Our study could help advertisers and marketers figure out how best to use these platforms for efficient and well-targeted marketing.

Acknowledgment

The authors would like to thank V V R Sastry, Vipin Tyagi and Ravi Gupta from C-DOT India for their valuable support and suggestions. This work was supported by the Commonwealth of Australia and the Department of Science and Technology, India, under the Australia-India Strategic Research Fund, and the Department of Information Technology, India under research project #RP02442.

9. REFERENCES

[1] S. Asur, B. A. Huberman, G. Szabo, and C. Wang. Trends in social media: Persistence and decay. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11, 2011.

[2] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. In *Proceedings of the 3rd international conference on Internet and network economics*, WINE '07, pages 306–311, San Diego, CA, USA, 2007. Springer-Verlag.

[3] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 665–674, Hyderabad, India, 2011. ACM.

[4] T. Carnes, C. Nagarajan, S. M. Wild, and A. van Zuylen. Maximizing influence in a competitive social network: a follower's perspective. In *Proceedings of the ninth international conference on Electronic commerce*, ICEC '07, pages 351–360, Minneapolis, MN, USA, 2007. ACM.

[5] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *Proceedings of the 3rd conference on Onsline Social Networks*, WOSN '10, 2010.

[6] R. Ghosh and K. Lerman. A framework for quantitative analysis of cascades on networks. In *Proceedings of the 4th ACM International Conference on Web search and data mining*, WSDM '11, 2011. Full version at http://arxiv.org/abs/1011.3571.

[7] J. L. Iribarren and E. Moro. Branching dynamics of viral information spreading. *Phys. Rev. E.*, To appear. http://arxiv.org/abs/1110.1884.

[8] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *Proceedings of the ACM Sigcomm workshop on Social Networks*, WOSN '08, 2008.

[9] 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*, WWW '10, pages 591–600. ACM, 2010.

[10] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 251–260. ACM, 2012.

[11] 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*, KDD '09, pages 497–506. ACM, 2009.

[12] G. Lotan. Data reveals that occupying twitter trending topics is harder than it looks! blog.socialflow.com, October 12 2011.

[13] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. *arXiv/1206.1331*, 2012.

[14] T. Rodrigues, F. Benvenuto, M. Cha, K. P. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *Proceedings of the 2011 Internet Measurement Conference*, IMC '11, 2011.

[15] 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. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 695–704, Hyderabad, India, 2011.

[16] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina. Correcting for missing data in information cascades. In *Proceedings of the 4th international conference on Web Search and Data Mining*, WSDM '11, pages 55–64, 2011.

[17] D. Sousa, L. Sarmento, and E. Mendes Rodrigues. Characterization of the twitter @replies network: are user ties social or topical? In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, pages 63–70, Toronto, ON, Canada, 2010. ACM.

[18] R. M. Tripathy, A. Bagchi, and S. Mehta. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1817–1820, Toronto, ON, Canada, 2010. ACM.

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

[20] S. Wu, C. Tan, J. Kleinberg, and M. Macy. Does bad news go away faster? In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 646–649, 2011.

[21] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 177–186, Hong Kong, China, 2011. ACM.

[22] S. Yardi and D. Boyd. Tweeting from the town square: Measuring geographic local networks. In *Proceedings of the fourth International AAAI Conference of Weblogs and Social Media*. The AAAI Press, 2010.