Modeling Twitter Engagement in Real-World Events

Yuheng Hu[1] and Yili Kevin Hong[2]
[1]University of Illinois at Chicago
[2]Arizona State University
yuhenghu@uic.edu, hong@asu.edu

Abstract

Twitter offers tremendous opportunities for people to engage with real-world events (e.g., political election) through information sharing and communicating about these events. However, little is understood about the factors that affect people's Twitter engagement (e.g., posting) in such real-world events. This paper examines multiple predictive factors associated with four different perspectives of users' Twitter engagement, and quantify their potential influence on predicting the (i) presence; and (ii) degree of the user's engagement with real-world events. We find that the measures of people's prior Twitter activities, topical interests, geolocation, and social network structures are all variously correlated to their engagement with real-world events.

1. Introduction

Twitter has emerged as one of the most important platforms for people to report, share, and communicate with others about various types of real-world events. These range from widely-known events (e.g., the U.S Presidential Debate) to smaller scale, local events (e.g., a local parade, or a car accident). Social media has many advantages over the traditional media channels, such as ubiquity, immediacy, and seamless communication in covering real-world events. Given these advantages, tweets can typically reflect events as they happen, in real-time. For this reason, recent years have witnessed a growing interest in research that aims to develop tools for real-world event detection and characterization based on social media posts [1], [2].

Unfortunately, little is understood thus far about the factors that affect people's engagement with real-world events on social media (e.g. posting or exchanging event-related tweets): *Does a person post tweets about an event because they are interested in the topics pertaining to that event? Are they instead engaged because their friends are also posting tweets about it? Or is their engagement a reflection of the fact that this is a local event?* Answering these questions holds the key to developing a wide range of applications such as advertising (e.g., a brand can target its potential customers more effectively through understanding of Twitter engagement in its product release event).

This work aims to explore predictive variables and quantify their influence on predicting a person's presence and degree of Twitter engagement with various real-world events. Specifically, the presence of a person's Twitter engagement in response to an event can be defined as the existence of at least one tweet (or RT or mention) that references that event. And the degree of the person's Twitter engagement is measured by the number of tweets that they post regarding that event; more such tweets indicate that they are more engaged with that event. We collect factors that could potentially affect a person's Twitter engagement in real-world events from four categories: (i) Twitter activities, (ii) tweets' content, (iii) the person's geolocation, and (iv) the person's social network structure. We construct two statistical models based on logistic and linear regressions to assess the relative contributions of these variables towards predicting the presence of a person's Twitter engagement and the degree of that Twitter engagement in real-world events. Our study reveals several insights about the presence and degree of Twitter users' engagement in real-world events are revealed. For example, in terms of the presence of engagement, we find that among all the predictive factors, a user's prior Twitter activity and her social network most significantly impact the presence of the user's engagement with events.

2. Background

Twitter and Real-World Events: As social media has become prominent in daily life, the evolving ways in which information is generated, viewed, and shared have inevitably transformed people's engagement with events [3]. Recent years have witnessed a growing research interest in developing tools for event identification and detection on social media [1], [4]. In addition, recent research also focuses on making sense of tweets and people's tweeting behavior around various real-

URI: http://hdl.handle.net/10125/41265
ISBN: 978-0-9981331-0-2

world events such as political events [5], [6], local events [7], and natural hazard events [8], [9].

Despite the rich literature on Twitter and its role in covering real-world events, to date, we are aware of little research that directly addresses the issue studied in this paper. The most relevant related work is on modeling predictive factors on social media for various other issues such as tie formation [10], tie break-up [11], tie strength [12] and retweeting [13]. Our effort differs from this past work in that we are exploring factors that may affect people's Twitter engagement in response to real-world events. Below, we discuss some background showing how a person's prior Twitter activities (e.g., communicating with others), her tweets' content (e.g., topical interests, linguistic styles), her geographical location, and her social networks relate to her Twitter engagement with real-world events.

Social Activity, Social Capital and Event Engagement: There are a lot of works in social science studying the performance of individuals and collectives to networks of social relationships. There are two strands that are most relevant to this work: One popular strand of the literatures focuses on how individuals use the network resources to achieve personal goals [37], and another strand relevant focuses on the utility of networks for collective endeavors, such as engagement with civic events or participation in political groups [14]. Both research strands deal with different aspects of social capital, defined as a collection of resources that either an individual or an organization can access through a set of communal norms, networks, and sanctions [14].

With the advent of social media, many researchers have focused on how social capital is fostered by social media and how such social activities by individuals and collectives affect social capital and their event engagement or political participation (as compared to previously studied *offline* social activities) [37]. Many works have successfully identified several kinds of individual or collectives' online social activities that affect social capital. These include directed communications with targeted individuals (e.g., Facebook private messages; Twitter replies, mentions, and favorites), broadcast communications which are not targeted at anyone in particular (e.g., Facebook wall updates), and passive consumption of content [17]. Moreover, the volume of social media posts (e.g., total number of tweets in a period) and the posting rate have also been shown to influence social capital [18]. On the other hand, many research found that the online social activities can still affect social capital and civic engagement [35, 36]. In particular, in a seminal work by Zúñiga et al, they found that seeking information via social network sites is a positive and significant predictor of people's social capital and civic participatory behaviors, online and offline.

Since previous work on online social activity and social capital mostly focus on political events. It is not clear the impact of social activity for more general events. Here, in this work, we empirically test whether a person's social activities help in predicting their engagement with different types of real-world events.

Twitter User Types, Topical Interests and Event Engagement: The "endurability" theory [19] shows that people are likely to remember a good experience and are willing to repeat it. Application of this theory here indicates that a person may be more likely to engage with an event if the topics related to that event are the same as – or at least similar to – the topic that the person is interested in on Twitter. There are many ways to infer a person's topical interests on Twitter, such as based on the content of the person's previous tweets [20] or the person's following list [21]. This is because, according to the principle of homophily, the similarity between individuals leads to a greater potential for interpersonal connections; when establishing connections, people tend to build relationships with others who are like them [22]. Sharing interests with another person is one form of similarity [23] that can be used to build relationships; this can lead to the follow relationship being established.

Geolocation and Event Engagement: It is known that a person's geographical location significantly affects their social connections and activities in the offline world. Recent research has also found evidence to show that offline geography has a significant impact on user interactions, tie formation, and information diffusion on online social media like Twitter [25]. In particular, researchers have discovered that users preferentially connect and exchange information with other users from their own country, and lesser information is exchanged across national boundaries. However, even such transnational links and interactions occur between users in geographically and linguistically proximal countries within their network. Similarly, researchers also identified that geographical proximity plays a key role in trend/innovation adoption [26]. Based on these results, we posit that a person's geolocation may affect their engagement with real-world events on Twitter if that person's location is geographically proximate to the event's location (e.g., a user may only care about events that happen in their neighborhood).

Social Networks and Event Engagement: The correlation between social network influence (e.g., net- work size and social ties) and user engagement has been studied extensively. For example, [27] showed that the relationship between online and offline network size and people's engagement with civic events is positive. They further found that network structure and social ties (especially weak ties) are determined to be strong predictors of the engagement. There are many different ways to form ties on Twitter, and ties can be formed either directly or indirectly. For example, following a person on Twitter can be seen as a direct tie. In such cases, dyadic properties such as reciprocity play key roles in the process of tie formation. On the other hand, ties can be formed indirectly such as through common network neighbors (known as transitive ties). For example, consider the case where three people form an undirected network: A and C are both friends of B, but A and C are not friends. However, as the number of common neighbors (occurrences of B) between A and C increases, the likelihood of an A-C tie being formed and the corresponding tie strength also increase [28]. Here we explore the extent to which these network sizes and tie formations impact a person's engagement in events as compared to the person's Twitter activities, topical interests, and geolocation information.

3. Data Collection and Set up

Obtaining Real-world Events and Events' Geolocations: To obtain real-world events, one possible approach is to first obtain an event list from newspapers and then get the corresponding tweets. However, such an approach is not applicable for several reasons. First, not every event reported by newspapers is popular/trending on Twitter. As [29] pointed out, the popularity of tweets is affected by multiple reasons aside from newsworthiness. Second and more importantly, such an approach will be significantly biased towards larger, more broadly newsworthy events due to the nature of newspapers, which could potentially misguide our analysis. To avoid this, we use a different approach by first detecting real-world events from Twitter streams, and then inferring their geolocations later. For the first step, we adopt the event detection framework [4] to automatically detect events from Twitter.

Next, to infer the geolocations of the real-word event clusters, we asked annotators to individually read a sample tweet from each real-world event's cluster to gain an understanding of what the event is really about. The annotators were then asked to find the geolocation of the event cluster via search engines by coming up with their own search keywords (e.g., event-related hashtags, timestamps) based on their event understanding. The annotation yields satisfactory results for use in this work (see below for more details).

Obtaining Twitter Users' Geolocations: To infer a Twitter user's present geolocation, we use two approaches. First, we infer the location directly from the user's event-related tweets if the present location is mentioned/attached. Otherwise, we infer the home location from the user's profile and tweets using the methods mentioned in [32] to infer the geolocations of Twitter users. We then verify the extracted location information with the diurnal patterns of the user's tweets [33]. For example, most people in New York City will tweet about having dinner and the nightlife between 5:00PM EST to 1:00AM EST. So if a person regularly posts tweets about lunch around 12:00AM EST, they probably are not from the New York City area. Based on our preliminary testing, we found this algorithm together with the diurnal pattern verification yielded stable performance (78.4% for cities).

Constructing the dataset: In practice, we first obtained nearly 37M English tweets from the Twitter firehose during August of 2014. We then applied the event detection algorithm (see [4]) on these tweets to find real-world events. As a result, we obtained 7,468 real-world event clusters. Next, we needed to infer the geolocations of these event clusters. We hired 20 annotators to read 10 sample tweets from each of their assigned event clusters (each annotator was assigned roughly 373 event clusters) and infer the geolocation. As a result of this step, our annotators were able to infer the geolocations (on city level) for 643 event clusters verified by two extra annotators who did not participate in the previous event geolocations inference task (with inner-rate k=0.76). Among these 643 event clusters with geolocations identified, 425 events happened in U.S (e.g., New York City, NY, Beverly Hills, CA, Ferguson, MO), and the rest were in Europe, the Middle East, and Asia. Finally, based on those 643 events, we obtained a total number of 22,957 Twitter users who posted at least three tweets in response to one of these events. We applied the location inference algorithm (see above) to predict the location (on city level) of each user. Besides, in order to calculate the measures for the predictor variables, we collected all the tweets posted by each user in the most recent six months preceding their first ever event engagement with any of the 643 events used.

4. Methods

In this section, we first present the dependent variables used in our predictive models, followed by a description of the predictor variables.

Dependent Variables:

1) Presence of a person's Twitter engagement in a real-world event: A binary measure that indicates

whether or not a person posts, replies to, or retweets tweets in relation to a particular event on Twitter

2) Degree of a person's Twitter engagement in a real-world event: A continuous measure that indicates the number of tweets that the person generates (via post, reply to, or retweet) relating to the event.

Predictor Variables: The literature reviewed in the previous section pointed us to the major kinds of predictor categories: Twitter activities, tweets' topics, geolocation, and social network structure. Following this, we collected predictors that are frequently used in the literature to reflect or affect a user's activities on Twitter, social capital, user types, geolocation, topics, and social network structure.

Variables related to Twitter activities
Total number of tweets. The total number of tweets a person has posted, reposted or replied.

Directed communications. The number of tweets with "@" plus the number of favorite tweets divided by the total number of tweets. This measure indicates interpersonal activities between the person and other users.

Broadcast communications. The ratio of tweets with no "@" at all in the tweet to total number of tweets in a period.

Ratio of retweets. The total number of times a person reposts other Twitter users' tweets, relative to the total number of tweets produced by the person in a period. This measure indicates how often the person interacts with other Twitter users and broadcasts those users' tweets to their own social circle (i.e., their followers).

Hashtag usage. This is defined as the ratio of tweets that contain at least one hashtag to the total number of tweets from a person in a period.

Meformer. This is computed as the ratio of meformer (who share tweets about themselves [33]) tweets to the total number of tweets by a person in a period. Following the approach used in [33], if a tweet contains any of the 24 self-referencing pronouns (e.g., "I", "me", "we", "us"), then it is classified as a meformer tweet. We also discard those tweets which contains both self-reference and third-person pronouns.

Informer. This is computed as the ratio of informer (who seek and share informational content) tweets to the total number of tweets by a person in a period. We identified informer tweets as those containing any of the 20 third-person pronouns (e.g., "He", "She", "it", "them"). In addition, if a tweet contains either a URL, "RT", "MT", or "via", we deem it an informer tweet as well.

Variables related to tweets' content
Topical interests from tweets' content. This measure is calculated as the topical similarity between two topic distributions: the first is computed based on a person's tweets in a period, while the second is computed based on all the event-related tweets (from other users) posted prior to the person's engagement with that event. In practice, assume a person *u* has posted T_u tweets in the past three months of the detected events in August 2014. Now, assume an event starts at 8:00PM and *u* engages with this event on Twitter (i.e., user u posts their first event-related tweet) at 8:30PM. Additionally, between 8:00PM and 8:30PM, there are T_Q event-related tweets posted by a set of other users *Q*. We then apply topic model LDA (we set the number of topics K = 20 in practice) [30] on both T_u and T_Q to learn the topic distributions respectively. We then measure the topical interest similarity between the two learned distributions based on JS-divergence. Intuitively, higher similarly indicates that the person's prior exhibited topical interests (reflected from their prior tweets' content) are closer to the event's topics (which are inferred from other people's event-related tweets).

Topical interests from the person's following list. This measure is calculated based on the topical similarity between the topics of the tweets written by the people that a person follows, and the event's topics. Unlike the way for inferring the topical interests based on person *u*'s recent tweets T_u, we following methods mentioned in [21]: first, given the following list of *u*, we obtain the 200 most recent tweets from each user on that list (due to Twitter API limitation). Next, we distill topic distributions from these tweets using LDA (we set the number of topics K = 20 in practice). We then compare these topics with the topics of T_Q (see above) to measure the topical interests similarity between the two learned distributions based on JS-divergence.

Variables related to geolocation information
Geographical proximity. This considers the geographical proximity between a person's location and the event's location. As indicated in the previous section, the dataset used in this study only includes Twitter users and events whose geolocations could be identified.

Variables related to network structure
Number of followings. The number of Twitter users that a person was following.

	Not engaged with events		**Engaged with events**		
Measure	Mean	SD	Mean	SD	Diff.
Event count	-	-	13.1	7.9	-
Tweet count			4.5	0.4	-
Twitter activity					
Total tweets	282.3	202.2	32.2	5.8	***
Direct communication	3.2	6.3	1.9	5.8	n/s
Broadcast Communication	1.1	0.8	1.8	1.2	***
Hashtag ratio	0.5	0.3	0.6	0.1	**
RT ratio	0.2	0.4	0.4	0.1	n/s
Meformer	0.4	0.1	0.3	0.2	*
Informer	0.4	0.2	0.5	0.4	*
Twitter content					
Topical interests From tweets content	-	-	0.25	0.1	-
Topical interests From following	-	-	0.2	0.1	-
Geolocation					
Geographical proximity	-	-	308mi	200.2	-
Social network					
Followers	403.2	150.1	435.2	112.2	n/s
Friends	210.4	226.2	180.4	101.2	n/s
Friends posted prior	-	-	4.33	5.08	-
Avg. common neighbor	-	-	9.3	10.4	-
News friends	5.9	8.2	7.2	7.1	**

Table 1. Mean and SD values for Twitter users' event engagement, compared to averaged values of these Twitter users' non-event tweeting behavior, and paired sample t-tests for the difference.

Number of followers. The number of Twitter users who were following the person.

Followings posted prior. The number of a person's followings who had already posted event-related tweets before the person posted to that event. As discussed earlier, since following (e.g., A follows B) forms a directed tie, it is possible that the person will be influenced to post tweets when a lot of their followings post about an event prior to their own engagement.

Average common neighbor prior. This measure examines the overlaps between the followings of a person a and the followings of user b, where a has already engaged in the event on Twitter while b has not. In the context of Twitter, a person's following list often represents their interests. Therefore, the common neighbor factor essentially measures the shared interests between two people. According to triadic closure, such a measure also indicates the tie strength between A and B [34].

Number of followings about news. This measure is defined as the total number of a person's followings who are deeply involved in news. To identify such news related accounts, we first obtain Twitter profiles for all of the person's friends. We then look at each profile to check which ones contain news related keywords such as "news", "reporter", "journalist", "TV" and so on. We deem those user's news related accounts. One motivation for this measure is that news agencies are often authorities and first-hand resources for reporting events. It is possible that if a person followers a lot of news agency accounts, then they will likely be interested in knowing about and engaging with real-world events.

5. Results

In the following section, we first provide descriptive statistics for the variables used in our statistical models. Following this, we present the contribution of these variables in predicting the presence and degree of people's Twitter engagement with real-world events.

5.1 Descriptive Statistics

Table 1 shows descriptive statistics (mean, standard deviation) for the number of events that one person engages with, and the event-related tweets that person posts – along with predictor variables – based on the event data we collected in August 2014. For comparison, we also generate statistics based on an event participant's regular tweets five months prior the the events (i.e., March 2014 to July 2014). We calculate the significance of the difference between these two situations. Note that some of the predictor variables are compared pair-wise, such as topical interests, geographic proximity and so on. Therefore, we only report the pair-wise statistics for the event data.

Also note that we excluded users that were extreme outliers (z-score > 4.0) with respect to our metrics for Twitter activity levels (e.g., the total number of tweets, maximum tweets per hour, etc) and follower/following counts. As a result, we removed 787 "outlier" users from a total of 22,957 users in our dataset, resulting a total of 22,170 people who have posted tweets over the course of 643 events. Within these messages, 28% of the messages had hashtags, 48% retweets, 27% direct replies, 33% links, and 68% mentions, indicating that the event participants were highly interactive.

Twitter activities: On average, a user engaged in 13.1 events over a month, and they posted 4.5 tweets per event. The Broadcast Communication shows the average number of tweets that are not directed to any specific person. During events, this rate is significantly higher. Such changes are also reflected in directed communication. The ratios of retweets and hashtag usage to the total number of tweets in a period are moderate for the majority of users retweets comprised about 15% of users' messages, and hashtags were used in about 20% of tweets. Compared to these, we witness significant changes during events – where the ratio of hashtags and retweets increases to 20% and 100% respectively. Combining these discoveries, we conclude that people tend to communicate more with others during an event that they are engaged in, thus showing a deeper involvement and engagement with the topics related to that event.

Besides, nearly half of users’ regular tweets are identified as “meformer” (40%), and the “informer” category accounts for 40% of tweets. However, in the context of event engagement, the percentage value of “informer” tweets witnessed an increase, and “meformer” tweets a decrease. This could indicate that people tend to hare more information (e.g., through retweets) during the course of an event. However, people do also continue posting information about their thoughts and their presence during the event.

Tweet content: In general, users show a fairly diverse range of topics that they post in relation to, which is reflected in and manifested as the relatively low topical similarity to actual event topics.

Geolocation: In terms of the geographical proximity between the event participants’ location and the event’s location, we found that most events were non-local to the event participants – this is reflected in that measure’s relatively high value (i.e., 308 miles between the inferred event participants’ locations and the events’ locations).

Social network: The majority of users have an average of 403 followers, and 150 friends. About 4.33 event participants who joined in the event prior to the target user’s engagement are the followings of that user. Moreover, for the people who posted prior to the target user but are not part of the following set, it is seen that there are around 10 common friends between those users and the target user. This indicates that one-hop weak ties do exist between event participants. Later we will demonstrate the strength of these predictors.

5.2 Prediction of presence

We now turn to the core question examined in this study: *to what extent do the independent variables used predict the presence, and degree, of a person’s Twitter engagement with a real world event?*

In order to examine the relative impact of these variables, we first standardized the measures, and then examined whether they predicted a user’s participation/engagement using a repeated measure (643 trials, or events) logistic regression. The question of whether or not the user participated was modeled as a binary dependent variable. Table 2 shows the results of this regression. An immediate insight that can be gleaned is that the total tweets posted by a user prior to her event engagement is a significant predictor of whether the user will take up or engage with an event. Specifically, as far as communication oriented tweets are concerned, both directed and broadcast communication are good indicators, albeit in opposite senses.

	Beta	**Standard Err.**	**Sig.**
Twitter activity			
Total tweets	0.15	0.05	***
Directed communication	-0.22	0.07	*
Broadcast communication	-0.01	0.00	***
Hashtag ratio	0.11	0.08	
RT ratio	0.49	0.09	*
Meformer	-0.11	0.1	
Informer	0.21	0.02	***
Tweet content			
Topical interests from tweet content	0.11	0.08	
Topical interests from following	0.12	0.09	
Geolocation			
Geographical Proximity	0.02	0.01	
Social network			
Followers	-0.04	0.02	*
Friends	-0.11	0.04	**
Friends posted prior	0.02	0.01	**
Avg. common neighbor prior	-0.2	0.1	**
News friends	0.13	0.02	**

Table 2. Prediction of presence: Logistic. Standardized variables in simultaneous repeated measures logistic regression predicting participation in events over 643 “trials”. Pseudo R2 = 0.37

The coefficients for those variables seem to indicate respectively that lower directed communication or higher broadcast communication correlate directly with higher engagement. This is fairly intuitive, since directed communication tends to be among a user’s friends and about non-event topics, and in most cases can only be seen by the mentioned users; while broadcast communication is intended for a wider audience consisting of all of the user’s followers. Finally, both the ratio of hashtags used and the ratio of retweets are positive indicators of event engagement; this is easy to see since RTs and hashtags respectively are two key ways in which a user can signal their active interest and affiliation with an event.

As far as the tweet content variables and Twitter user types variables are concerned, we did not find evidence of the topical interests being good predictors of engagement with events that display those same topics. However, we will show later in our analysis that when the tweets are broken down by topic and not considered as a single monolithic set, these topic-specific correlations become stronger predictors of engagement. As regards meformer versus informer tweets, the meformer tweets are not very good predictors of engagement, which is obvious since such tweets mostly involve the user talking about things that are highly personalized and hyperlocal to their own lives. Informer tweets, on the other hand, display a positive correlation to engagement; since such tweets are usually in the third person, this result combined with the broadcast communication considered previously indicate that a user who posts such tweets will usually engage with something that multiple other users are also interested in (hence an event as against a personalized happening).

As concerns geolocation, we did not find any significant evidence – in contrast with prior research [25] – that the geographical proximity has any effect on a user's engagement with an event. This would seem to indicate that users will choose to engage with an event whether or not it is "local" (in their surrounding vicinity) or non-local.

Finally, where the social network variables are concerned, we find that all of the variables are predictors with at least some degree of significance (and some more so than others). Interestingly, the only positive correlation is with the number of new friends. A further manual inspection revealed that most of the news friends' posts actually are occurring before the user starts contributing messages and engaging with the event. This indicates that users are inspired and motivated to engage with events when they see tweets from news agencies relating to those events on their timelines. However, this only goes so far – as the negatively correlated variables show, a large number of friends/followers and neighbors may bring down awareness, engagement, and subsequent participation (i.e., their coefficients are negative). We argue that this can be possibly attributed to a variety of factors. Some of these may include cognitive overload on the part of the target user, higher noise, posts being perceived as less personal, and most importantly, a perception that the topic is already sufficiently covered, e.g., posted by friends (thus reducing an "informer" user's motivation in engaging with it).

5.3 Prediction of Degree

To further explore the relative impact of these variables in predicting the degree of prediction in new events, we performed a linear regression, using participation levels in past events to predict the level of participation in a final, target event. The results are shown in Table 3.

We find that the most significant predictors of the degree of a user's engagement happen to be the social network variables, followed by the twitter activity variables. Specifically, the only social network variable that shows a significant positive correlation is the number of posts from the user's friends prior to the user's engagement with the event, which can be explained in terms of the activity that a user sees on their timeline with regard to that event. However, as in the previous case, increases in the user's network size seem to dampen the degree of engagement somewhat (which can be attributed to many of the same reasons described previously). The participant's own past tweet content seemed to have no significant effect on the predicted degree of engagement, save for the total and broadcast tweets, which offer a historical window into how active the user was in general.

	Beta	**Standard Err.**	**Sig.**
Twitter activity			
Total tweets	0.37	0.05	***
Directed communication	-0.1	0.07	*
Broadcast communication	0.04	0.00	***
Hashtag ratio	0.09	0.01	***
RT ratio	0.069	0.09	*
Meformer	-0.06	0.1	
Informer	0.02	0.02	***
Tweet content			
Topical interests from tweet content	0.12	0.08	
Topical interests from following	0.07	0.03	
Geolocation			
Geographical Proximity	0.01	0.01	
Social network			
Followers	-0.04	0.02	*
Friends	-0.07	0.02	**
Friends posted prior	-0.02	0.01	**
Avg. common neighbor prior	-0.22	0.09	**
News friends	0.13	0.02	**

Table 3. Prediction of degree: OLS coefficients for standardized variables in simultaneous repeated measures logistic regression predicting participation in events over 643 "trials". Adjusted R2 = 0.56

6. Discussion, Implications, and Limitations

At the beginning of this paper, we posed five important questions relating to the engagement of users on social media with real-world events; and whether such engagement (and its level) could be effectively and practically predicted based on information available from that social media. In this section, we consider possible answers to those questions that are suggested by the data and revisit the related theories to examine our answers.

Does a person post tweets about an event because they are interested in the topics pertaining to that event?

Our analysis confirms that this is indeed the case. To highlight this, we point the reader to the analysis concerning prediction of presence and degree (Table 2 and 3), and the contrast with the similar prediction analysis given a breakdown of the events into different topics (Table 4). In the former case, there is no significant indicator of correlation from the content of a user's tweets to their engagement with an event. However, in the latter, there is a marked increase in the significance of the correlation between the content of tweets related to events in specific topics, and the user's engagement with those events (e.g., politics & business, tech & science, and sports). This is exactly what the "endurability" theory [19] proves: people are likely to remember a good experience and are willing to repeat it. In other words, people like to repeatedly talk about the topics that they are most familiar with/interested in. So, they will show deeper engagement in those specific topics, in contrast to boarder and more general topics.

Are they instead engaged because their friends are also posting tweets about it?

The answer to this is positive as well, conditioned on the type of event that the user is engaging with. We have shown in the previous section that certain kinds of events – local events, as well as odd news – users tend to engage more due to their friends (following list) posting content relating to those events prior to the user's own engagement. This verifies the discoveries by Zuniga et al. [27] network structure and social ties (especially weak ties) are determined to be strong predictors of the civic engagement. We also extend their theory by discovering the social network and time affects on the engagement with real-world events (indeed, some events are about civic issues).

Perhaps they are just a very active user of Twitter?

The degree to which a user was active on Twitter (the number of tweets posted by them) does indeed show a strong correlation across all cases to their predicted engagement with an event. This correlation seems to be agnostic of the type of event (as against the previous two questions, above), and hence it seems likely that more active users are more likely to be interested and engaged in a new event, across the board. This finding validates our earlier conjecture that these activities will first directly affect people's engagement in events on social media; such engagement will later indirectly affect social capital. Our finding extends existing ffffflit-eratures on the relationship between social media activities and social capital [17], [16] by exploring the role of user engagement.

Is their engagement a reflection of the fact that this is a local event? The answer to this question reverts to the pattern of dependence on the kind of event observed in the answers to the first two questions. There are certain kinds of events that can be classified as engaging to a user primarily due to their local nature – as described in the previous section, these tend to be sports and local events. The connection to local events is obvious and trivial; a user in New York City is unlikely by and large to care about events that are happening in (say) far-off Tulsa, Oklahoma. For sports, it is likely that users within a given geographical area are more likely to care about teams that call that particular area home (although of course there will always be outliers; however, our analysis is focused on the typical user).

Limitations: Although the data that we use and the results produced from that data seem to imply some rather strong conclusions, certain limitations of the study must also be considered when going forward. The first of these is the categorization of events: although the categories we use in this study are quite general, and capture a large portion of the posts on Twitter, arguments can certainly be made in support of finer-grained categories that will support more nuanced analysis with respect to users' potential engagement with events. Additionally, the event detection and classification process that is currently used by us can be further improved – both to classify events better, and to allot events across different categories (as against just a single category, as is the case currently). We also did not consider people's personality in the study. It is possible that certain personality (e.g., openness and extraversion) may affect people's event engagement. Besides, the setup of locality did not consider the difference between big cities and small cities. It is possible that within a big city (e.g., New York City), within 100 miles may still be considered as "local". Furthermore, we did not consider the contextual factors such as the nature and timing of real-life events which may affect the engagement. Finally, in this study, we did not consider the fact that there may exist different kinds of target users when engagement with events is under consideration. While we did partition a target user's following list coarsely (in terms of friends, news accounts, etc.), the target users themselves may also be distributed across various categories that exhibit some correlation (and hence predictive power) with respect to event engagement.

7. Conclusion

In this paper, we developed statistical models of people's Twitter engagement with real-world events. Categories of engagement predictors were conceptually developed, operationalized, and assessed for their relative impact on users' engagement presence, and the degree of that engagement. We explored the relative impact of multiple measures collected from four different user perspectives: prior Twitter activity, tweets' content, geolocation or geographic proximity, and social network structure. In particular, we found several key factors that predict the users' presence in engagement with real-world events, including total number of tweets, communication modes, friends' engagement in events, etc. We also examined the effects of these predictors in predicting the degree of engagement. We also examined the effects of these factors with respect to the different types of events predicated on their topics. We concluded that users' prior activities, as well as their social network structure, can be very good predictors for both the presence and the degree of their engagement with real-world events. Given a finer granularity of events (according to their topics), the content of tweets and the geographic proximity provide additional predictive power with respect to different event categories.

8. References

[1] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 851–860.

[2] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources," in Pro- ceedings of the first SIGMM workshop on Social media. ACM, 2009, pp. 3–10.

[3] 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. ACM, 2010, pp. 591–600

[4] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real- world event identification on twitter." ICWSM, vol. 11, pp. 438–441, 2011.

[5] Y. Hu, A. John, D. D. Seligmann, and F. Wang, "What were the tweets about? topical associations between public events and twitter feeds." in ICWSM, 2012.

[6] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in VAST, 2010

[7] Y. Hu, S. D. Farnham, and A. Monroy-Herna´ndez, "Whoo. ly: Facilitat- ing information seeking for hyperlocal communities using social media," in CHI 2013

[8] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," in CHI 2010.

[9] K. Starbird, "Delivering patients to sacre´ coeur: collective intelligence in digital volunteer communities," in CHI 2010

[10] S. A. Golder and S. Yardi, "Structural predictors of tie formation in twitter: Transitivity and mutuality," in Social Computing (SocialCom) 2010

[11] F. Kivran-Swaine, P. Govindan, and M. Naaman, "The impact of network structure on breaking ties in online social networks: unfollowing on twitter," in CHI'11

[12] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in CHI 2009

[13] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in Socialcom 2010

[14] B. Wellman and S. Wortley, "Different strokes from different folks: Community ties and social support," American journal of Sociology, vol. 96, no. 3, p. 558, 1990.

[15] D. V. Shah, "Civic engagement, interpersonal trust, and television use: An individual-level assessment of social capital," Political Psychology, vol. 19, no. 3, pp. 469–496, 1998.

[16] J. B. Hyman, "Exploring social capital and civic engagement to create a framework for community building," Applied Developmental Science, 2002.

[17] M. Burke, R. Kraut, and C. Marlow, "Social capital on facebook: Dif- ferentiating uses and users," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2011, pp. 571–580.

[18] C. Hutto, S. Yardi, and E. Gilbert, "A longitudinal study of follow predictors on twitter," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2013, pp. 821–830.

[19] J. Read, S. MacFarlane, and C. Casey, "Endurability, engagement and expectations: Measuring children's fun," in Interaction Design and Children, 2002

[20] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, “Analyzing user modeling on twitter for personalized news recommendations,” in UMAP 2011

[21] M. Burgess, A. Mazzia, E. Adar, M. Cafarella, “Leveraging noisy lists for social feed ranking,” in ICWSM

[22] M. McPherson, L. Smith-Lovin, and J. M. Cook, “Birds of a feather: Homophily in social networks,” Annual review of sociology, pp. 415– 444, 2001.

[23] S. L. Feld, “The focused organization of social ties,” American journal of sociology, pp. 1015–1035, 1981.

[24] M. Naaman, J. Boase, and C.-H. Lai, “Is it really about me?: message content in social awareness streams,” in Proceedings of the 2010 ACM conference on Computer supported cooperative work. ACM, 2010, pp. 189–192.

[25] J. Kulshrestha, F. Kooti, A. Nikravesh, and P. K. Gummadi, “Geographic dissection of the twitter network.” in ICWSM, 2012.

[26] J. L. Toole, M. Cha, and M. C. Gonza´lez, “Modeling the adoption of innovations in the presence of geographic and media influences,” PloS one, vol. 7, no. 1, p. e29528, 2012.

[27] H. G. de Zu´ n˜ iga and S. Valenzuela, “The mediating path to a stronger citizenship: Online and offline networks, weak ties, and civic engage- ment,” Communication Research, vol. 38, no. 3, pp. 397–421, 2011.

[28] D. Cartwright and F. Harary, “Structural balance: a generalization of heider’s theory.” Psychological review, vol. 63, no. 5, p. 277, 1956.

[29] L. Hong, O. Dan, and B. D. Davison, “Predicting popular messages in twitter,” in WWW 2011, pp

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” JMLR 2003.

[31] Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: a content-based approach to geo-locating twitter users,” in CIKM, 2010, pp. 759–768.

[32] J. Mahmud, J. Nichols, and C. Drews, “Home location identification of twitter users,” TIST, 2014.

[33] M. Naaman, A. X. Zhang, S. Brody, and G. Lotan, “On the study of diurnal urban routines on twitter.” in ICWSM, 2012.

[34] S. Wasserman, Social network analysis: Methods and applications. Cambridge university press, 1994

[35] Gil de Zúñiga, H., Jung, N., & Valenzuela, S. (2012). Social media use for news and individuals' social capital, civic engagement and political participation. Journal of Computer-Mediated Communication, 17(3), 319-336.

[36] S. Valenzuela. Unpacking the use of social media for protest behavior the roles of information, opinion expression, and activism. American Behavioral Scientist, 57(7), 920-942.

[37] Erickson, Bonnie H. "Culture, class, and connections." American journal of Sociology (1996): 217-251.