$1.00 per RT #BostonMarathon #PrayForBoston: Analyzing Fake Content on Twitter

Aditi Gupta*, Hemank Lamba**, Ponnurangam Kumaraguru*
*Indraprastha Institute of Information Technology, Delhi, India
**IBM Research Labs, Delhi, India
Email: {aditig, pk}@iiitd.ac.in, helamba1@in.ibm.com

***Abstract*—Online social media has emerged as one of the prominent channels for dissemination of information during real world events. Malicious content is posted online during events, which can result in damage, chaos and monetary losses in the real world. We analyzed one such media i.e. Twitter, for content generated during the event of Boston Marathon Blasts, that occurred on April, 15th, 2013. A lot of fake content and malicious profiles originated on Twitter network during this event. The aim of this work is to perform in-depth characterization of what factors influenced in malicious content and profiles becoming viral. Our results showed that 29% of the most viral content on Twitter, during the Boston crisis were rumors and fake content; while 51% was generic opinions and comments; and rest was true information. We found that large number of users with high social reputation and verified accounts were responsible for spreading the fake content. Next, we used regression prediction model, to verify that, overall impact of all users who propagate the fake content at a given time, can be used to estimate the growth of that content in future. Many malicious accounts were created on Twitter during the Boston event, that were later suspended by Twitter. We identified over six thousand such user profiles, we observed that the creation of such profiles surged considerably right after the blasts occurred. We identified closed community structure and star formation in the interaction network of these suspended profiles amongst themselves.**

I. Introduction

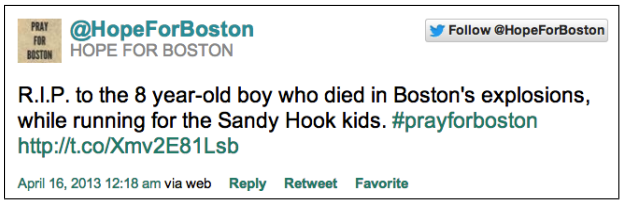
Emergence of online social media (OSM) and their increasing popularity, has created a new medium and arena for e-crime. Online social media provides people with an open platform to share information and opinions on diverse topics. Twitter is a micro-blogging service, which has gained popularity as one of the prominent news source and information dissemination agent over last few years [24]. During real-world events like earthquakes, elections and social movements, we see a sudden rise in activity over the Internet [25]. People log-on to Twitter and other social media, to check for updates about these events, and to share information and opinions about the event. Twitter can act like an effective crowd-sourced crisis management medium used by general public and authorities to coordinate efforts during crisis events. The content on Twitter can provide rich information about an event, however, this vast resource of information is often is not credible, true and full of noise [23]. Various forms of e-crimes like spam, phishing, spreading rumors, fake information and identity theft engulf the social media experience [5] [15]. Hence there is a dire need to study the adverse effects of these e-crimes on ground to real people.

During crisis and emergency events, due to heightened anxiety and emotional vulnerabilities, people are often more susceptible to fall for rumors / fake content. In one of the most recent incidents in U.S.A., Dow Jones index plunged 140 points due to a rumor tweet posted from a news agency's (Associated Press) Twitter account [14]. the estimated temporary loss of market cap in the S&P 500 totaled $136.5 billion. The rumor mentioned that U.S.A. president Barack Obama has been injured in twin explosions at the White House. In case of *England Riots*, social media was responsible for spreading and instigating violence amongst people. Many rumors propagated during the riots, which resulted in large scale panic and chaos among the public [34]. Two people were also sentenced for spreading false posts on Facebook during the riots [10]. In another incident in Venezuela, some people had spread rumors on Twitter, to destabilize the banking system of the country [44]. In one of the extreme case, *Twitter terrorists* in Mexico were given thirty years sentence for spreading rumors about a fake shooting by gunmen in schools [1]. As parents rushed to get their children from school, and all telephone lines were jammed, the incorrect information, caused a massive chaos in the city, 26 road accidents, jammed telephone lines and created chaos among people. In case of Boston marathon blasts, the rumors, resulted in fake and fraud charity fund creations, and incorrect news about a young boy dying in the blasts. Figure 1 presents some such sample tweets of rumors / fake content during Boston blasts. In another incident in India, social media was used during ongoing communal riots to spread inflammatory and provoking false content against the government [38].

The aim of this paper is to characterize and propose solutions to counter various forms of malicious activities on Twitter during events such as the Boston blasts. In this paper we used data collected during the Boston blasts, for the analysis done in this paper. We collected about 7.8 million tweets for the Boston marathon blasts using the Twitter APIs. Our data collection was limited from the fact that it was started 45 minutes after the blasts had occurred. To the best of our knowledge this is one of the largest studies, to analyze a dataset of tweets containing fake information / rumors. Also, this work presents the first comprehensive characterization of content posted on Twitter during the Boston blasts, with special focus on fake content propagation. In future, we aim to validate our results and approach during other kinds of crisis events, like natural disasters and political protests.

The territory of social media for e-crimes is challenging since anonymity, private settings and large volume of data present challenges for researchers to build affective solutions. OSM have a short impact time, i.e., the millions of users

(a)



(b)

Fig. 1. Two sample tweets containing fake content. (a) A tweet from a fake charity profile. (b) Rumor about a child being killed in the blasts.

get affected by an e-crime on social media such as Twitter and Facebook, within a few hours [14]. Hence, the solutions built need to work in real-time and be capable of handling large volume and evolving characteristics. The three main research questions we aim to explore in our work are: Firstly, characterize the user attributes of people who propagate fake content during crisis events. Can simple user attributes like number of followers and account being verified, be used to differentiate between fake and true news? Secondly, to evaluate, if impact of users who propagate fake content be used to estimate how viral the content would become in future? Thirdly, what kind of interactions occur between the suspended accounts on Twitter, that are created during a crisis event?

Boston Marathon Blasts: Twin blasts occurred during the Boston Marathon on April 15th, 2013 at 18:50 GMT. Three people were killed and 264 were injured in the incident [37]. Two suspects Tamerlan Tsarnaev (deceased) and Dzhokhar Tsarnaev (in custody) carried out the bombings. There was a huge volume of content posted on social media websites, including Twitter, after the blasts. We saw online social media being effectively used by Boston Police to track down the suspects and pass on important information to the public. There were various malicious entities which spread false information and posted fake content. To name a few specific cases: tweets about fake charities, offering to donate money to Boston victims became highly viral; rumor about some children who were running the marathon died in the blasts, along with fake pictures of them were circulated. Figure 2 shows a picture clicked during the blasts. [1] Timeline of social media coverage of Boston blasts has been analyzed and visualized by some people. [2]

There were two primary kinds of fake content that emerged on Twitter during the Boston marathon blasts. We present analysis about the domain of fake information creation and

[1]Taken image from http://www.telegraph.co.uk/news/worldnews/northamerica/usa/9996332/Boston-Marathon-explosions-three-dead-dozens-injured-as-bombs-hit-race-finish-line.html

[2]http://beta.seen.co/event/boston-marathon-2013-boston-ma-2013-7033

Fig. 2. A picture clicked during the Boston marathon blasts.

propagation, along with suspended profiles on Twitter during crisis events. Our main contributions are:

- We characterized the spread of fake content on Twitter using temporal, source and user attributes. We found that approx. 75% of fake tweets are propagated via mobile phone devices.
- We applied linear regression model to predict how viral fake content would in future based on the attributes and impact of users currently propagating the content.
- We analyzed the activity and interaction graphs for the suspended user profiles created during Boston blasts. We identified that malicious user exploit the event happening to indulge in e-crimes like impersonation and propaganda spreading.

This paper is organized as follows: Section II discusses the literature review about the problem domain of analyzing malicious content on Twitter. Section III describes the methodology and description of work done in this research work. Section IV summarizes the temporal, spatial, impact analysis for the propagation of fake content. Section V presents the network, text and user attributes from suspended profiles created during Boston blasts. Section VI contains the discussion and future work.

II. Related Work

A. Role of OSM during Real World Events

Role of social media has been analyzed by computer scientists, psychologists and sociologists for impact in the real-world. Palen et al. presented a vision on how Internet resources (technology and crowd based) can be used for support and assistance during mass emergencies and disasters [30]. They also studied two real world events, to understand and characterize the wide scale interaction on social networking websites with respect to the events [31]. Sakaki et al. used tweets as social sensors to detect earthquake events [35]. They developed a probabilistic spatio-temporal model for predicting the center and trajectory of an event using Kalman and particle filtering techniques. Based on the above models, they created an earthquake reporting application for Japan, which detected the earthquake occurrences based on tweets and sent user's, alert emails. In another research work, Sakaki et al. analyzed tweet

trends to extract events that happened during a crisis event from Twitter [36]. They analyzed log of user activity from Japanese tweets on all earthquakes during 2010-2011. Cheong et al. performed social network analysis on Twitter data during Australian floods of 2011 to identify active players and their effectiveness in disseminating critical information [11].

Work has been done to extract situational awareness information from the vast amount of data posted on OSM during real-world events. Vieweg et al. analyzed the Twitter logs for the Oklahoma Grassfires (April 2009) and the Red River Floods (March and April 2009) for presence of situational awareness content. An automated framework to enhance situational awareness during emergency situations was developed. They extracted geo-location and location-referencing information from users' tweets; which helped in increasing situational awareness during emergency events [40]. Verma et al. used natural language techniques to build an automated classifier to detect messages on Twitter that may contribute to situational awareness [39]. Another closely related work was done by Oh et al., where they analyzed Twitter stream during the 2008 Mumbai terrorist attacks [29]. Their analysis showed how information available on online social media during the attacks aided the terrorists in their decision making by increasing the terrorist's *social awareness*. Corvey et al. analyzed one of the aspects of applying computational techniques and algorithms to social media data to obtain useful information for social media content, i.e. linguistic and behavioral annotations [16]. One important conclusion obtained by them was that during emergency situations, users use a specific vocabulary to convey tactical information on Twitter.

Mendoza et al. used the data from 2010 earthquake in Chile to explore the behavior of Twitter users for emergency response activity [25]. The results showed that propagation of rumor tweets versus true news were different and automated classification techniques can be used to identify rumors. Longueville et al. analyzed Twitter feeds during forest Marseille fire event in France; their results showed that in location based social networks, spatial temporal data can be analyzed to provide useful localized information about the event [17]. A team at National ICT Australia Ltd. (NICTA) has been working on developing a focused search engine for Twitter and Facebook that can be used in humanitarian crisis situation [18]. Hughes et al. in their work compared the properties of tweets and users during an emergency to normal situations [2]. They performed empirical and statistical analysis on the data collected during disaster events and showed an increase in the use of URLs in tweets and a decrease in @-mentions during emergency situations.

B. Assessing Quality of Information on OSM

Presence of spam, compromised accounts, malware, and phishing attacks are major concerns with respect to the quality of information on Twitter. Techniques to filter out spam / phishing on Twitter have been studied and various effective solutions have been proposed. Chhabra et al. highlighted the role of URL shortener services like *bit.ly* in spreading phishing; their results showed that URL shorteners are used for not only saving space but also hiding the identity of the phishing links [12]. In a followup study Aggarwal et al. further analyzed and identified features that indicate phishing tweets [4]. Using URL, domain, network and user based features, they detected phishing tweets with an accuracy of 92.52%. One of the major contributions of their work, was the Chrome Extension they developed and deployed for real-time phishing detection on Twitter. Grier et al. characterized spam spread on Twitter via URLs. They found that 8% of 25 million URLs posted on Twitter point to phishing, malware, and scams listed on popular blacklists [21]. Ghosh et al. characterized social farming on Twitter, and also proposed a methodology to combat link farming [20]. Yang et al. analyzed community or ecosystem of cyber criminals and their supporters on Twitter [42]. Yardi et al. applied machine learning techniques to identify spammers [43] and obtained 91% accuracy. They used features: (1) searches for URLs; (2) username pattern matches; and, (3) keyword detection. Benevenuto et al. classified real YouTube users, as spammers, promoters, and legitimates [6]. They used techniques such as supervised machine learning algorithms to detect promoters and spammers; they achieved higher accuracy for detecting promotors; the algorithms were less effective for detecting spammers. Nazir et al. provided insightful characterization of phantom profiles for gaming applications on Facebook [27]. They proposed a classification framework using SVM classifier for detecting phantom profiles of users from real profiles based on certain social network related features.

Now, we discuss some research work done to assess, characterize, analyze and compute trust and credibility of content on online social media. Truthy [7], was developed by Ratkiewicz et al. to study information diffusion on Twitter and compute a trustworthiness score for a public stream of micro-blogging updates related to an event to detect political smears, astroturfing, misinformation, and other forms of social pollution [33]. It works on real-time Twitter data with three months of data history. Castillo et al. showed that automated classification techniques can be used to detect news topics from conversational topics and assessed their credibility based on various Twitter features [9]. They achieved a precision and recall of 70-80% using J48 decision tree classification algorithms. Canini et al. analyzed usage of automated ranking strategies to measure credibility of sources of information on Twitter for any given topic [8]. The authors define a credible information source as one which has trust and domain expertise associated with it.

Gupta et al. in their work on analyzing tweets posted during the terrorist bomb blasts in Mumbai (India, 2011), showed that majority of sources of information are unknown and with low Twitter reputation (less number of followers) [23]. This highlights the difficulty in measuring credibility of information and the need to develop automated mechanisms to assess credibility of information on Twitter. The authors in a follow up study applied machine learning algorithms (SVM Rank) and information retrieval techniques (relevance feedback) to assess credibility of content on Twitter [22]. They analyzed fourteen high impact events of 2011; their results showed that on average 30% of total tweets posted about an event contained situational information about the event while 14% was spam. Only 17% of the total tweets posted about the event, contained situational awareness information that was credible. Another, very similar work to the above was done by Xia et al. on tweets generated during the England riots of 2011 [41]. They used a supervised method of Bayesian Network to predict the credibility of tweets in emergency situations. Donovan et al focused

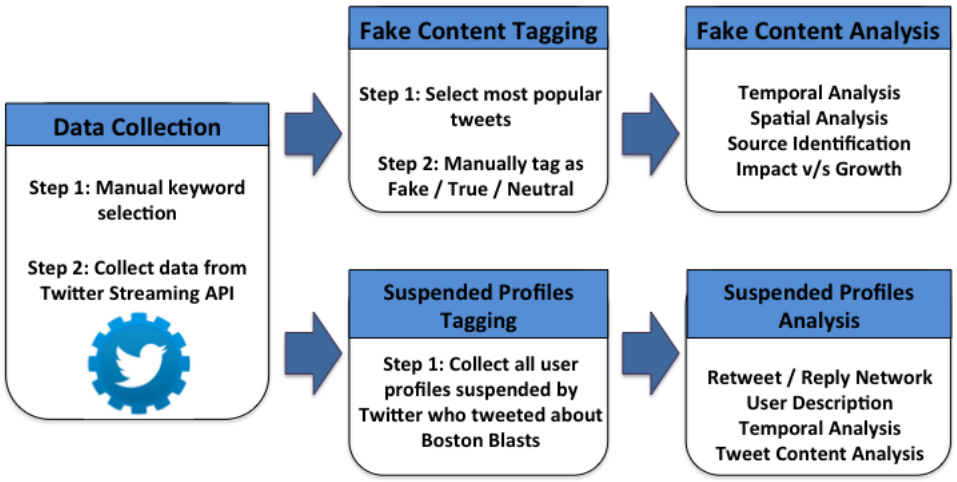


Fig. 3. Architecture diagram describing the methodology followed in this paper for analyzing fake content and suspended profiles on Twitter during the Boston marathon blasts.

their work on finding indicators of credibility during different situations (8 separate event tweets) were considered [28]. Their results showed that the best indicators of credibility were URLs, mentions, retweets and tweet length. Qazvinian et al. prepared a log-likelihood based retrieval model which used content based, network based and Twitter based features to extract misinformation (rumor) tweets and misinformers (people who propagate rumors) from Twitter [32]. Nagy et al. introduced Credo, a semi-supervised system which predicts the credibility of the messages on Twitter. They use a model based on PageRank to come up with a credibility score for each of the post. [3]

A different methodology, than the above papers was followed by Morris et al., who conducted a survey to understand users perceptions regarding credibility of content on Twitter [26]. They asked about 200 participants to mark what they consider are indicators of credibility of content and users on Twitter. They found that the prominent features based on which users judge credibility are features visible at a glance, like username and profile picture of a user. Another approach to detect users with high value users of credibility and trustworthiness was taken by Ghosh et al., they identified the topic based experts on Twitter [19]. Using techniques based on the wisdom of the Twitter crowds; they used the Twitter Lists feature to identify experts in various topics.

A lot of research work has been done on analyzing various forms of e-crimes on online social media. Similarly, a lot of researchers have analyzed content generated on OSM during real world events, but not a lot of researchers have worked on the intersection of these two problems, particularly in context of *identifying fake content during real world events*. In this paper, we collected a large dataset of fake tweets during the Boston blasts, and presented an in-depth characterization and analysis of propagation of rumors on Twitter.

III. Methodology

In this section, we discuss the data collection and annotation methodology in detail. Figure 3 presents the architecture diagram of the methodology followed in this paper. We describe the methodology to characterize and analyze the landscape of malicious content posted on Twitter during real world events. We identified malicious tweets by tagging the most viral tweets during the events as fake or true news. To identify malicious accounts, we selected the user accounts that were created just after the Boston blasts and were later suspended by Twitter for violating its terms and conditions.

A. Data Collection

Twitter provides an interface via its APIs to enable researchers and developers to collect data. The three APIs provided by Twitter are namely *REST, STREAMING and SEARCH* APIs. Streaming API is used to get tweets and their corresponding user's data in real time, satisfying some user specified filtering (based on keywords, location, language, etc.). We used the Streaming API of Twitter to collect tweets related to Boston blasts [38]. We used the following keywords to collect data: *Dzhokhar Tsarnaev, #watertown, #manhunt, Sean Collier, #BostonStrong, #bostonbombing, #oneboston, bostonmarathon, #prayforboston, boston marathon, #bostonblasts, boston blasts, boston terrorist, boston explosions, bostonhelp, boston suspect*. We collected about 7.9 million unique tweets by 3.7 million unique users. The descriptive statistics of the data are given in Table I. Our data collection was started about

TABLE I. Descriptive statistics of dataset for Boston blasts, April 15th - 19th, 2013.

Total tweets	7,888,374
Total users	3,677,531
Tweets with URLs	3,420,228
Tweets with Geo-tag	62,629
Retweets	4,464,201
Replies	260,627
Time of the blast	Mon Apr 15 18:50 2013
Time of first tweet	Mon Apr 15 18:53:47 2013
Time of first image of blast	Mon Apr 15 18:54:06 2013
Time of last tweet	Thu Apr 25 01:23:39 2013

45 minutes after the blast. But since many tweets of the first 45 mins, got retweeted later, we were also able to capture those

in our data collection mechanism. This is the largest known dataset of Boston marathon blasts. Within 3 minutes of the blasts happening, we got our first tweet; and within 4 minutes of the blasts the first picture of the blast was posted on Twitter, which we captured in our dataset.

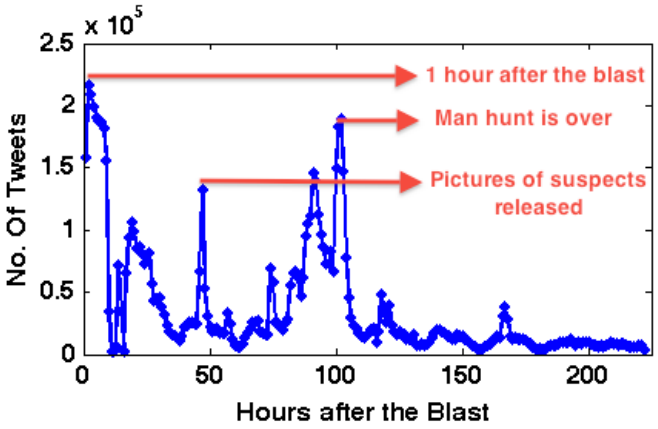


Fig. 4. Timeline for the tweets collected for the Boston marathon blasts.

Figure 4 shows the temporal distribution of the tweets collected for the blasts. We have annotated the figure, with the corresponding real world happenings to understand when the activity on Twitter peaked. Boston blasts and the manhunt of suspects was an event that generated a lot of posts on Twitter. Many people offered help and coordinated relief measures via Twitter. The Boston police used its official account *boston_police* to spread the photograph of the suspects and got aid in their manhunt.

In all 0.8% [62,629 / 7,888,374] of total tweets during the Boston blasts, shared geo-location in their tweets. Figure 5 shows the distribution of the geo-tagged tweets. On Twitter retweets done using the retweet button do not have geo-location field. For the 500,00 annotated tweets considered by us in this paper which were retweets of the top twenty most viral tweets, did not have any geo-location information in them.

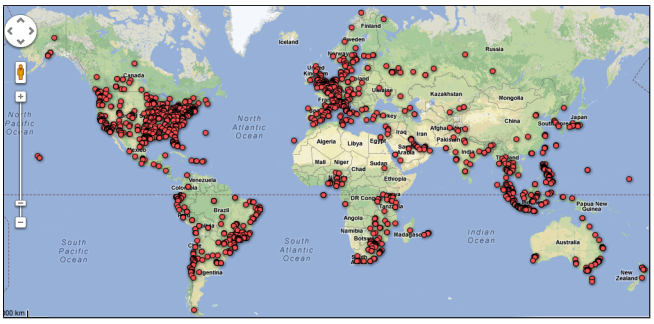


Fig. 5. Geo-location for the tweets collected for the Boston marathon blasts. We observe that such impactful events draw posts from all over the globe.

B. Annotated Data

We identified the top 20 most popular tweets during the Boston Marathon blasts. In total, the top 20 tweets constituted 453,954 total tweets (6% of all 7.9 million Boston tweets). We manually tagged these tweets in three categories: True , Rumor / Fake and NA. NA stands for *Not Applicable*, and it represents tweets which neither give any true or fake information about the event, they are mostly personal opinions and condolences. Table II shows the tweets, their number of retweets and their corresponding tag. We found that 51% of the tweets were generic comments and opinions of people, with neither true or fake information. The percentage of fake tweets was much more (29%) than true information (20%).

There were 3,249 overlap in users who tweeted both true information and fake, 3% of 94,383 unique users who tweeted true information and 2% of 147,169 unique users who tweeted fake information tweets. Although, the time period of fake and NA category tweets were quite overlapping, we found only a overlap of 2,895 users in both the categories. These observations imply, that each set of users who tweeted fake / true / NA category of tweets are unique from each other. Since we considered only retweets and replies to most viral tweets, which were retweeted by the user using retweet button, we can be reassured that all retweets carried the same exact text and hence also belonged to the same category.

IV. Characterizing Fake Content Propagation

In this section, we analyze various aspects of spread of fake information tweets and propagation during the Boston marathon blasts.

A. Temporal Analysis

To analyze the temporal distribution of the tweets posted during the Boston blasts, we calculate the number of tweets posted per hour. Figure 6 shows the cumulative growth of the three categories of tweets over five days. We plotted the log values on the Y-axis, to avoid bias from the large amount of total number of tweets. The inset graph shows the growth of the data for the first fifty hours only. For the first 7-8 hours, we mostly observe only tweets belonging to the NA and fake category were being spread in the network. The

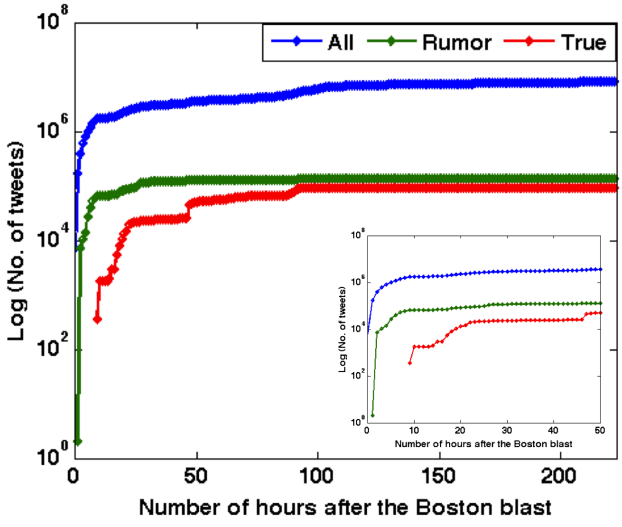


Fig. 6. The log distribution for the number of the total, fake, true information and NA category tweets. The inset figure presents the results for the first 50 hours after the blast.

circulation of true information only starts after about eight hours from the time of the blasts. After a few hours only,

TABLE II. TOP 20 MOST POPULAR TWEETS (RETWEETED AND REPLIED). WE TAGGED EACH OF THE TWEETS TO BELONG TO THE FOLLOWING THREE CATEGORIES: FAKE / RUMOR , TRUE AND NOT APPLICABLE (NA). ABOUT 51% OF THE MOST VIRAL TWEETS BELONGED TO NA CATEGORY, I.E. CONSISTING OF COMMENTS AND OPINIONS OF PEOPLE.

RTs	Tweet Text	Category
87,903	#PrayForBoston	NA
33,661	R.I.P. to the 8 year-old girl who died in Boston's explosions, while running for the Sandy Hook kids. #prayforboston http://t.co/WhaaTG3nSP	Fake / Rumor
30,735	Dzhokhar Tsarnaev, I have bad news for you. We never lose at Hide and Seek, just ask Bin Laden and Saddam. Good Luck.Sincerely, America	NA
28,350	For each RT this gets, $1 will be donated to the victims of the Boston Marathon Explosions. #DonateToBoston	Fake / Rumor
27,163	#prayforboston	NA
26,954	Reports of Marathon Runners that crossed finish line and continued to run to Mass General Hospital to give blood to victims #PrayforBoston	Fake / Rumor
26,884	In our time of rejoicing, let us not forget the families of Martin Richard, Lingzi Lu, Krystle Campbell and Officer Sean Collier.	True
20,183	I will DONATE $100 for EVERY pass I catch next season to whatever Boston Marathon Relief Fund there is. And $200 for any dropped pass.	True
18,727	Doctors: bombs contained pellets, shrapnel and nails that hit victims #BostonMarathon @NBC6	True
17,673	#prayforBoston	NA
17,560	For every retweet I will donate 2 to the Boston marathon tragedy! R.I.P!	Fake / Rumor
16,457	From Sarasota to Boston, our thoughts go to the victims of the marathon bombings. We're saddened by loss of life and injuries to so many....	NA
13,788	So far this week- #prayfortexas - #prayforboston - two 14 year olds killed a homeless man as a dare- bomb threats It's only Thursday	True
13,610	Jhar #manhunt @J_tsar. Look at this from a follower. Look at the time if the tweet http://t.co/xgnAJpeVTr	NA
13,482	BREAKING: Suspect #1 in Boston Marathon bombing shot and killed by police. Suspect #2 on the run, massive manhunt underway.	True
13,275	#prayforboston	NA
12,354	BREAKING: An arrest has been made in the Boston Marathon bombings, CNN reports.	True
12,209	R.I.P. to the 8 year-old boy who died in Boston's explosions, while running for the Sandy Hook kids. #prayforboston http://t.co/Xmv2E81Lsb	Fake / Rumor
11,950	For each RETWEET this gets, $1 will be donated to the victims of the Boston Marathon Bombing.	Fake / Rumor
11,036	#WANTED: Updated photo of 19 year-old Dzhokhar Tsarnaev released. Suspect considered armed & dangerous. http://t.co/pzps8ovJTb	True

official and news user profiles give out confirmed and new information, which becomes viral. Atleast for the initial hours after a crisis, we need to distinguish fake / rumor tweets from only the generic comments and opinions of the people. For fake category tweets, we see that the first hour has slow growth, but once it becomes viral they have a very steep growth. This may be attributed to the fact that the user profiles (source of a fake tweet) are people with low social status and unconfirmed identity. Hence the initially fake tweet spread is slow, and they become highly viral only after some users with high reach (for e.g. large number of followers) propagate them further.

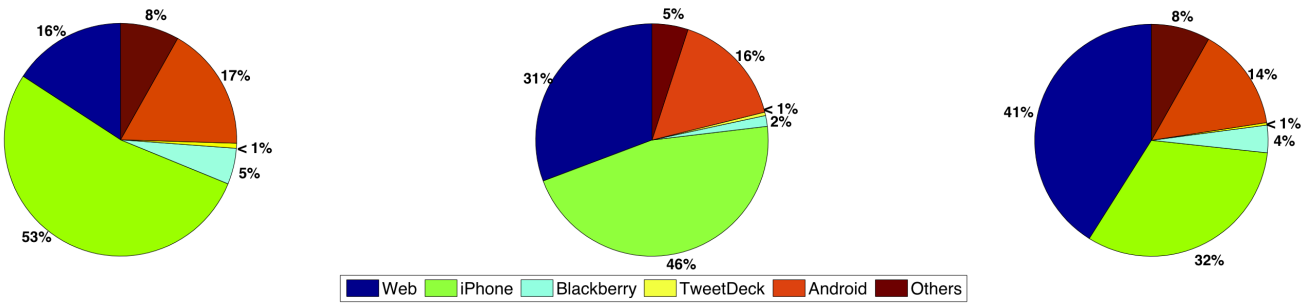
B. Fake Tweet Seed User Accounts

We analyzed the attributes and activities of the user accounts from where the fake tweets originated. Table III presents the various user profile attributes for the seed of the fake tweet user profiles. Of the six fake tweets identified, two users had started two rumors each. For most of the fake tweets we observe that the seed users are people with very few followers. *Seed 4* is the only user profile with high number of followers. The tweet posted by *seed 4* was *Reports of Marathon Runners that crossed finish line and continued to run to Mass General Hospital to give blood to victims #PrayforBoston*. This tweet even though was false and classified as fake content / media by the media too, [3] was harmless and not even deleted by Twitter. For all other sources, except *seed 4* we can say that the originators of the fake content are users with low credibility. We checked for the presence of these seed user profiles on Twitter now; all accounts except *seed 4* have been suspended by Twitter.

TABLE III. DESCRIPTIVE STATISTICS OF THE FOUR USER ACCOUNTS THAT WERE THE SEEDS OF THE SIX FAKE TWEETS.

	Seed 1	Seed 2	Seed 3	Seed 4
Number of Followers	10	297	249	73,657
Profile Creation Date	Mar 24 2013	Apr 15 2013	Feb 07 2013	Dec 04 2008
Number of Statuses	2	2	294	7,411
Number of Fake Tweets	2	2	1	1
Current Status	Suspended	Suspended	Suspended	Active

[3] http://www.guardian.co.uk/commentisfree/2013/apr/16/boston-marathon-explosions-reveal-twitter



(a) (b) (c)

Fig. 7. Tweet source analysis for the three categories of tweets: (a) Fake (b) True (c) NA. We observed that in case of fake tweets, approx. 75% users use mobile devices to tweet, as compared to 64% for true and 50% for NA category of tweets.

C. Tweet Source Analysis

We studied the source of the tweets that were posted. We analyzed the medium through which the tweets were posted. The results for the same are presented in Figure 7. We found that the tweets containing information (Fake or True) propagated more through mobile devices like iPhone, Android, Blackberry, etc. whereas the general non-informative tweets (NA category) were posted more via web interface. We found that approx. 75% of fake tweets are propagated via mobile phone devices, as compared to 64% true tweets and only 50% generic comments shared via mobile devices. This implies that people are eager to share the informative tweets and also willing to do that while being on the go. For non-informative tweets, people don't feel such urgency and post tweets mostly if they are accessing through the web interface.

D. Role of User Attributes in Fake Content Identification

To understand what kind of users aid in propagation of each category of tweets, we analyzed three main attributes of user profiles on Twitter. We calculated the average number of followers of the user accounts and the number of verified accounts that propagated the true, fake and NA tweets. Figure 8 summarizes the results for the first 120 hours after the blasts. We see that the average number of followers is the maximum for NA tweets, followed by true and fake tweets. Even though high number of users tweet generic news, the rumors get more viral. Number of people retweeting fake information tweets drops significantly in the latter hours (80-120 hours), this maybe so, as people start realizing that it is a rumor. We also observed that a high number of verified accounts propagate fake content, which is quite surprising. We can conclude that determining whether some information is true or fake, based on only factors based on high number of followers and verified accounts is not possible in the initial hours. The high number of verified and large follower base users propagating the fake information, can be considered as the reason for the fake tweets becoming so viral. It becomes difficult for the users to differentiate which sources to trust and which not. In the next section, we validate this hypothesis, by exploring if these user attributes can be used to estimate how viral a rumor / fake information tweet would become in future.

E. Role of User Attributes in Propagating Fake Content

We aim to understand, if user attributes can be used to estimate how viral fake content would become in future. Knowledge about how viral and damaging fake content can be in future can help us be prepared. In additional to basic user attributes like number of followers, friends, verified accounts, etc. we also define and compute an overall *impact* metric; to measure impact of users who propagate a tweet in making a tweet viral. We used user profile attributes to come up with a metric which calculates the overall *impact* of a user. We take the $impact$ of user as a linear combination of the following metrics:

- Social Reputation:- We take social reputation to be a function of the number of followers and the number of times the user has been listed. Number of followers denote the popularity of the users and number of times listed indicate how many people classified him in one of the custom list.

$$SocialReputation[SR(u_i)]$$

$$= \frac{log(n(fol))}{Max(log(n(fol)))} + \frac{n(listed)}{Max(n(listed))}$$

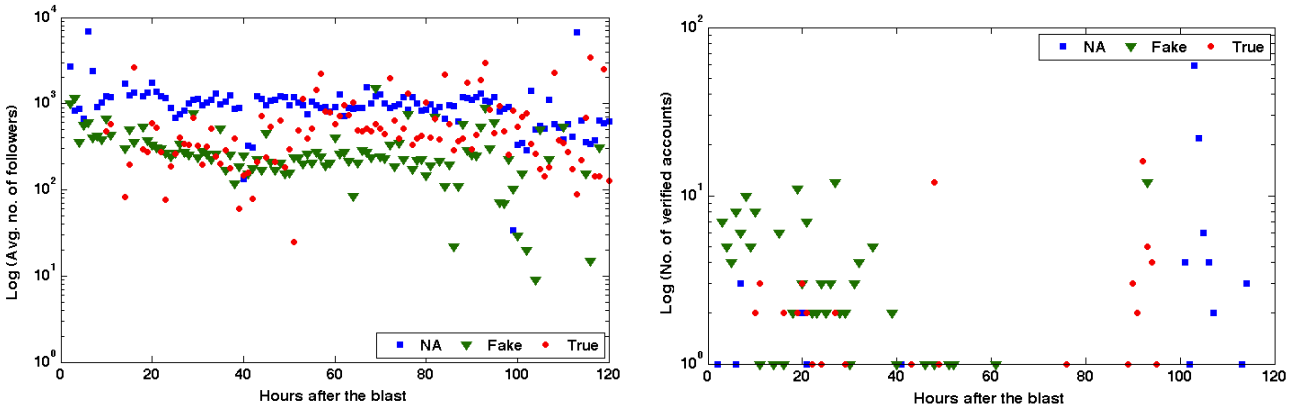
- Global Engagement:- It is how often does the user engage with posting activity on Twitter by tweeting, replying and retweeting. We take it as the ratio of the number of statuses the person has put to the time (in hours) since his account creation.

$$GlobalEngagement[GE(u_i)] = \frac{n(status)}{age}$$

- Topic Engagement:- We want to see how well a user is engaged in the current ongoing topic. For our context, the topic is the event under consideration. We measure this by number of tweets the user has posted on the particular topic

$$TopicEngagement[TE(u_i)] = \frac{n(status_t)}{max(n(status_t))}$$

- Likability:- The Likability of a user is to measure in general how much his content is liked by his followers or other users. We take it as the ratio of number of



(a) (b)

Fig. 8. Distribution of (a) average number of followers of the user accounts and (b) number of verified accounts in the annotated data.

times his statuses have been made favorite, to that of number of statuses posted.

$$Likability[L(u_i)] = \frac{n(favorited)}{n(status)}$$

- Credibility:- Credibility $C(u_i)$ of a news is based on how verifiable a user is. We take it to be 0, if the account is not verified by Twitter, else we take it as 1.

We define *impact* for a user, u_i, as a linear combination of the above mentioned components.

$$Impact[u_i] = SR(u_i) + GE(u_i) + TE(u_i) + L(u_i) + C(u_i)$$

Regression Analysis: We predict how the propagation will be in the immediate next time quantum. We used linear regression for this purpose. Our basic hypothesis is that $Impact$ of all the previously active users can be used to predict how many new users will get activated in the next time segment. For calculating the regression between attributes and growth of fake tweets, we consider time quantum of 30 minute each. For a particular time quantum, all users will have a similar contribution towards the cumulative $Impact$, so we weigh the cumulative impact according to the Poisson distribution.

$$CummulativeImpact(t) = \sum_{i=1}^{t-1} Impact(t-i) \times \exp^{(t-i)}$$

We estimate the number of people that are going to be activated in the next time segment using Linear Regression as follows:

$$N_{Active}(t+1) = \alpha \times CummulativeImpact(t) + \beta$$

For evaluation of linear regression, we used R^2 measure. The R^2 measure indicates with how much confidence can the model so created can account for the variability in the output data. Results of the model were compared with individual features as well and are presented in Figure 9. On an average for impact metric we achieve approx. 0.7 value of R^2. These results show us that it is possible to predict how viral a fake information tweet would become in future based on the attributes of the users currently propagating the fake content.

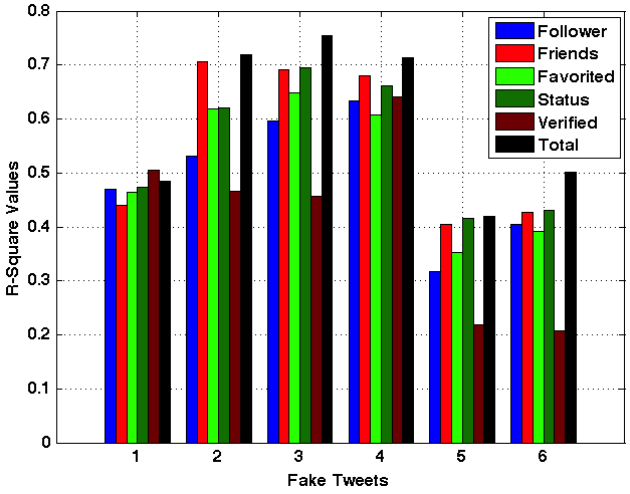


Fig. 9. Regression results of the overall impact of the users in previous time quantum. These results show us that it is possible to predict how viral the fake content would become in future based on the attributes of the users currently propagating the fake content.

V. Suspended Profiles Analysis

Hundreds of new accounts on Twitter get created everyday, many of these accounts are often malicious and spam accounts. In this section, we aim to identify the characteristics and activities of malicious new accounts created during the Boston marathon blasts. We identified 31,919 new Twitter accounts that were created during the Boston blasts tragedy [Apr. 15th - Apr. 20th], that also tweeted atleast one tweet about the event. Out of these 19% [6,073 accounts] were deleted or suspended by Twitter, when we checked two months after the blasts. Some of these accounts were quite influential during the Boston tragedy too. Next, we tried to find out how affective were these accounts during the Boston marathon events. Graph in Figure 10 shows the number of suspended profiles created in the hours after the blast. We observe that there are a lot of malicious profiles created just after the event occurs. Such profiles and accounts aim to exploit the high volume of content and interest of users in the event to spread spam, phishing and rumors. We constructed a network graph $G = (V, E)$ for the interaction between these newly created malicious profiles.

Where each node in V represents a suspended user profile, and an edge between two suspended nodes represents a retweet, reply or mention action by them.

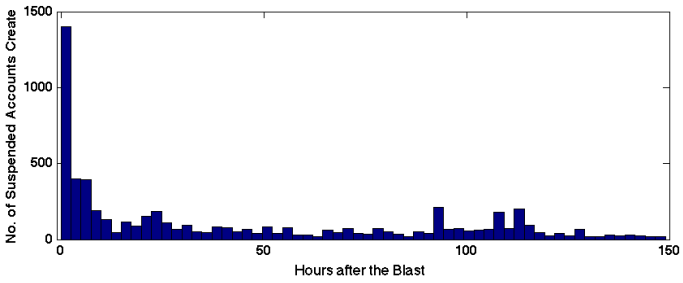


Fig. 10. This graph shows the number of suspended profiles created in the hours after the blast. We observe that there are a lot of malicious profiles created just after the event occurs.

Figure 11 shows the network obtained (some of the usernames are anonymized). We have removed all nodes with degree of zero, we found 69 nodes out of 6,000 suspended profiles had an edge to another node in the graph. Though the number of edges may look small, but we observe some interesting characteristics of the resulting network graph formed. We found four types of interactions amongst these accounts [left to right in Figure 11]:

- *Single Links:* We saw the content posted by a suspended user profile is propagated by one or two other suspended Twitter profiles. Some of these links are also bi-directional, indicating a mutual agreement between the nodes. This technique of creating multiple spam accounts to promote mutual content is often used by spammers on Twitter [20].
- *Closed Community:* We observed a community of users who retweet and mention each other, and form a closed community, as indicated by high closeness centrality values for the nodes. All these nodes have similar usernames too, all usernames have the same prefix and only numbers in the suffixes are different. This indicates that either these profiles were created by same or similar minded people for posting common propaganda posts. We then analyzed the text posted by these users. These twelve accounts were all tweeting the same propaganda and hate filled tweet. Since, Twitter does not allow users to post multiple posts with same content, another strategy applied by these accounts is tweeting the same text as above, but changing one character in each of the tweets. In all we found, 156 tweets by these 12 accounts.
- *Star Topology:* A fake account *BostonMarathons* was created similar to the original Twitter account *bostonmarathon*, resulting in users getting confused between the two, leading to a lot of propagation of content by the fake BostonMarathons profile. Impersonation or creating fake profiles is a crime that results in identity theft and is punishable by law in many countries.
- *Self Loops:* We saw that some of the profiles mentioned themselves in their tweets, resulting in self loops in the graph. This may be done by the users to self promote the content posted by them, as doing so brings them in the most recent tweets timeline of Twitter.

We saw that a large number of malicious accounts were created during crisis events. Next, amongst the suspended user profiles we searched for profile specifically created for exploiting the event. Some of them created related hoax profiles by using usernames similar to original accounts. We searched for the presence of the term *boston* in the name and username of the six thousand suspended profiles. We found 83 profiles which satisfied this criteria. Figure 12 shows the tag-cloud of the user description of these profiles. We found most of these profiles exploited the sympathy of people by using words such as *prayforboston, prayers, victims*. We can also see the malicious intent of people, as they attempt to create hoax accounts, as indicated by usage of words such as *official account*. The account *BostonMarathons* was also one such account which tried to impersonate the real *bostonmarathon* account.

VI. Discussion

Rumors or fake or incorrect information spread via online social media, have resulted in chaos and damage to people in the real world. Specially, during crisis events like earthquakes, bomb blasts and political uprisings, rumors can be very harmful. Malicious entities exploit the vulnerable emotions of people during crisis to make their rumors viral. Online social media, in particular, Twitter, is a mass media with reach to millions of users across the globe. Over recent years, misinformation on Twitter had resulted in damages ranging from financial to human lives. Detection and curbing of fake information on social media, is a relatively new and unexplored domain.

Our aim in this paper, was to characterize and provide useful insights into the domain of fake content propagation on Twitter. We collected about 7.8 million tweets for the Boston marathon blasts using the Twitter APIs. Our data collection was limited from the fact that it was started 45 minutes after the blasts had occurred. We analyzed source users of the fake tweets, spatial and temporal details of the fake tweets. We attempted to find out reasons that govern how viral (and in turn harmful) a fake information tweet becomes. We explored, using simple metrics, can we predict how the fake tweet would propagate in future. Another kind of fake content that is present on Twitter are the fake / spam user profiles. We analyzed six thousand malicious profiles that were created on Twitter right after the Boston blasts and were later suspended by Twitter.

Some of the major challenges in real time rumor detection and control on online social media are the following:

- *Volume of Content*: Most of the popular online social websites have users of the order of hundreds of millions. A huge amount of content is generated every second, minute and hour of the day. Any algorithms or solutions build to detect rumors on OSM should be scalable enough to process content and user data up to the order of millions and billions.
- *Short Impact Time*: Impact of malicious activities in online social media, such as, spread of spam, phishing

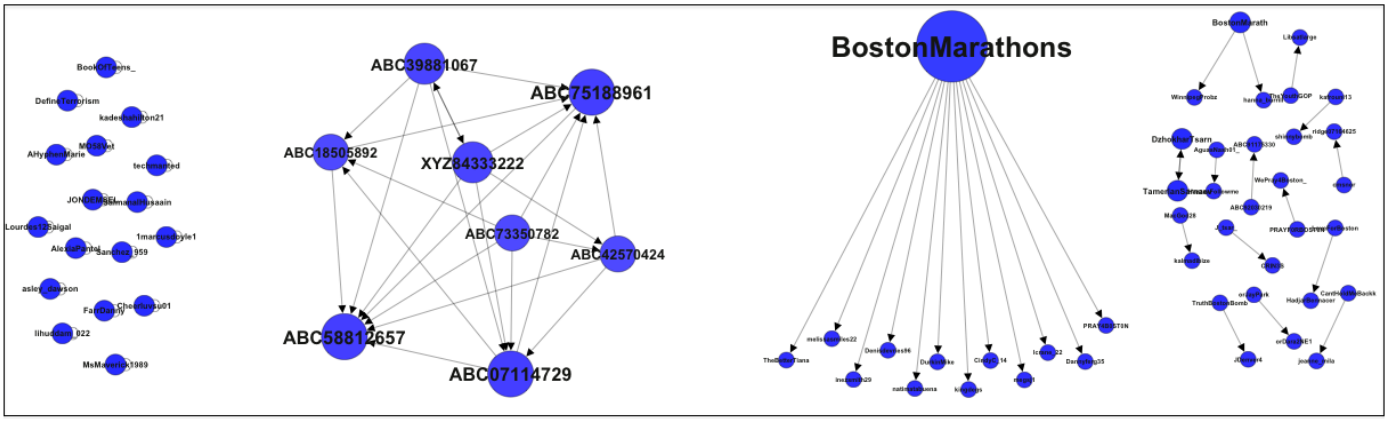


Fig. 11. Network of suspended accounts (retweets / replies / mentions) created during the Boston blasts. We see four different forms of interactions amongst the suspended profiles (left to right): Single Links, Closed Community, Star Topology and Self Loops.

(a) (b)

Fig. 12. Tag cloud of (a) user description and; (b) tweets posted by the suspended accounts created targeting specifically the Boston event. In the left tag cloud, we see the usage of words such as *official account* indicating profiles were created impersonating real profiles.

or fake content, causes vast amount of damage within hours of being introduced. Hence, solutions and algorithms built need to be able to solve and detect such content in real-time. Post-analysis may be able to capture concerned criminals, but would not be able to contain the damage.

- *Anonymity*: The Internet, by itself is an anonymous medium. Lack of tracing technologies, distributed framework and private regulations worsens the condition. People can create one or multiple profiles on OSM, with little or no verification required. Hence, validation and authenticity of any information source on OSM is extremely challenging.

We observed some differences in temporal propagation of rumors and true news. Since fake content tweets are often started by people with low number of followers on Twitter, they have a slow growth in the beginning, and then become viral as they are retweeted by higher influence people. While true news were mostly generated by accounts such as *boston_police* which already had 276,838 number of followers and hence, the true news have a more steady growth from the beginning. Spatial analysis did not give us much insights since the annotated tweets for the category of Fake / True / NA (Not Applicable) categories did not have much tweets with geo-tagged information. This was surprising, we found only 1 tweet out of approx. 500,000 tweets containing geo-location data. We found that approx. 75% of fake tweets are propagated via mobile phone devices. Next, we explored, if attributes of user profiles who propagate fake information can be used to answer the following two questions: Firstly, we identified fake tweets in the Twitter network using user profile attributes; and secondly, we used the attributes of user profiles to predict which rumor / fake tweet would become viral in the future on Twitter. For the first question, we observed that the user attributes are quite similar for fake, true news and NA category tweets, like a lot of verified and high number of followers accounts also retweet rumor tweets. Hence, user attributes by themselves are not very useful in identifying the fake tweets. For the second question on use of user attributes for predicting how viral fake content will get in future gave us more positive results. We used logistic regression to estimate the popularity of a fake tweet in future (30 mins to 1 hour in advance), based on the overall impact and other characteristics of the users who are currently propagating the fake tweets.

We saw a great number of malicious profiles created just after the Boston blasts, which were later suspended by Twitter. Out of the six thousand such profiles created (who tweeted about the event), we found about 83 profiles which further targeted the blasts, by creating usernames having the word 'boston' in it. On manual observation, we found many of these accounts to be fake accounts impersonating the real accounts (e.g. fake username: *bostonmarathons*, real username: *bostonmarathon*). The network analysis of the interactions among these suspended profiles revealed interesting techniques applied by them to spread their propaganda or spam; for e.g. we found a community of users who had similar usernames, and posted the same hatred targeting tweets multiple times. They also retweeted and replied to each other's tweets, to increase their visibility.

In future, we would like to verify the results we obtained for other real world events. We would like to evaluate if the features and modeled learned from the events in post-hoc can

be used to predict fake content and malicious profiles in real time. We are working towards building a real time technology to analyze the content generated on Twitter can be used to detect fake content and profiles is their early stages.

VII. ACKNOWLEDGMENTS

We would like to thank Government of India for funding this project. We would like to express our sincerest thanks to all members of Precog research group at IIIT, Delhi, for their continued support and feedback on the project. We would like to say special thanks to Anupama Aggarwal, who helped us in Twitter data collection for the Boston marathon blasts.

REFERENCES

[1] Jo Adetunji. 'twitter terrorists' face 30 years after being charged in mexico. *http://www.guardian.co.uk/world/2011/sep/04/twitter-terrorists-face-30-years*, 2011.

[2] Leysia Palen Amanda L. Hughes. Twitter Adoption and Use in Mass Convergence and Emergency Events. *ISCRAM Conference*, 2009.

[3] A. Antonucci, D. Huber, M. Zaffalon, P. Luginbuehl, I. Chapman, and R. Ladouceur. CREDO: a military decision-support system based on credal networks. In *Proceedings of the 16th Conference on Information Fusion (FUSION 2013)*, 2013.

[4] Ponnurangam Kumaraguru Anupama Aggarwal, Ashwin Rajadesingan. Phishari: Automatic realtime phishing detection on twitter. *7th IEEE APWG eCrime Researchers Summit (eCRS)*, 2012.

[5] APWG. Phishing activity trends report. 2013.

[6] Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 620–627, New York, NY, USA, 2009. ACM.

[7] Research Project by Indiana University. Truthy. *http://truthy.indiana.edu/*, 2011.

[8] Kevin R. Canini, Bongwon Suh, and Peter L. Pirolli. Finding credible information sources in social networks based on content and social structure. In *SocialCom*, 2011.

[9] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM.

[10] CBS. Brits get 4 years prison for facebook riot posts. *http://www.cbsnews.com/2100-202_162-20093364.html*, 2011.

[11] France Cheong and Christopher Cheong. Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods. In *PACIS*, 2011.

[12] Sidharth Chhabra, Anupama Aggarwal, Fabricio Benevenuto, and Ponnurangam Kumaraguru. Phi.sh/$ocial: the phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS '11, pages 92–101, New York, NY, USA, 2011. ACM.

[13] Taylor Clark. The 8 1/2 laws of rumor spread. *Psychology Today*, 2009.

[14] CNBC. False rumor of explosion at white house causes stocks to briefly plunge; ap confirms its twitter feed was hacked. *http://www.cnbc.com/id/100646197*, 2013.

[15] Symantec Corporation. Istr: internet security threat report 2013. 2013.

[16] William J. Corvey, Sudha Verma, Sarah Vieweg, Martha Palmer, and James H. Martin. Foundations of a multilayer annotation framework for twitter communications during crisis events. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[17] Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. "omg, from here, i can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, LBSN '09, pages 73–80, New York, NY, USA, 2009. ACM.

[18] ehealth. Crisis management: using twitter and facebook for the greater good. *http://leifhanlen.wordpress.com/2011/07/22/crisis-management-using-twitter-and-facebook-for-the-greater-good/*, 2011.

[19] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, 2012.

[20] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna PhaniGummadi. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, 2012.

[21] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, CCS '10, pages 27–37, New York, NY, USA, 2010. ACM.

[22] Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pages 2:2–2:8, New York, NY, USA, 2012. ACM.

[23] Aditi Gupta and Ponnurangam Kumaraguru. Twitter explodes with activity in mumbai blasts! a lifeline or an unmonitored daemon in the lurking? IIIT, Delhi, Technical report, IIITD-TR-2011-005, 2011.

[24] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue 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, New York, NY, USA, 2010. ACM.

[25] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 71–79, New York, NY, USA, 2010. ACM.

[26] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, New York, NY, USA, 2012. ACM.

[27] Atif Nazir, Saqib Raza, Chen-Nee Chuah, and Burkhard Schipper. Ghostbusting facebook: detecting and characterizing phantom profiles in online social gaming applications. In *Proceedings of the 3rd conference on Online social networks*, WOSN'10, 2010.

[28] J. O'Donovan, B. Kang, G. Meyer, T. Hllerer, and S. Adali. Credibility in context: An analysis of feature distributions in twitter. *ASE/IEEE International Conference on Social Computing, SocialCom*, 2012.

[29] Onook Oh, Manish Agrawal, and H. Raghav Rao. Information control and terrorism: Tracking the mumbai terrorist attack through twitter. *Information Systems Frontiers*, 13(1):33–43, March 2011.

[30] Leysia Palen, Kenneth M. Anderson, Gloria Mark, James Martin, Douglas Sicker, Martha Palmer, and Dirk Grunwald. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference*, ACM-BCS '10, 2010.

[31] Leysia Palen and Sarah Vieweg. The emergence of online widescale interaction in unexpected events: assistance, alliance & retreat. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, CSCW '08, pages 117–126, New York, NY, USA, 2008. ACM.

[32] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying Misinformation in Microblogs. 2011.

[33] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. WWW '11, 2011.

[34] Jonathan Richards and Paul Lewis. How twitter was used

to spread and knock down rumours during the riots. *http://www.guardian.co.uk/uk/2011/dec/07/how-twitter-spread-rumours-riots*, 2011.

[35] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[36] Takeshi Sakaki, Fujio Toriumi, and Yutaka Matsuo. Tweet trend analysis in an emergency situation. In *Proceedings of the Special Workshop on Internet and Disasters*, SWID '11, pages 3:1–3:8, New York, NY, USA, 2011. ACM.

[37] theTelegraph. Boston marathon blast injuries toll at 264. *http://www.dailytelegraph.com.au/boston-marathon-blast-injuries-toll-at-264/story-e6freuz9-1226628261632*, 2013.

[38] Twitter. Streaming api. *https://dev.twitter.com/docs/streaming-apis*, 2013.

[39] Sudha Verma, Sarah Vieweg, William Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.

[40] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1079–1088, New York, NY, USA, 2010. ACM.

[41] Xin Xia, Xiaohu Yang, Chao Wu, Shanping Li, and Linfeng Bao. Information credibility on twitter in emergency situation. In *Proceedings of the 2012 Pacific Asia conference on Intelligence and Security Informatics*, PAISI'12, 2012.

[42] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, 2012.

[43] Sarita Yardi, Daniel Romero, Grant Schoenebeck, and Danah Boyd. Detecting spam in a Twitter network. *First Monday*, 15(1), January 2010.

[44] Wendy Zeldin. Venezuela: Twitter users arrested on charges of spreading rumors. *http://www.loc.gov/lawweb/servlet/lloc_news?disp3_l205402106_text*, 2010.