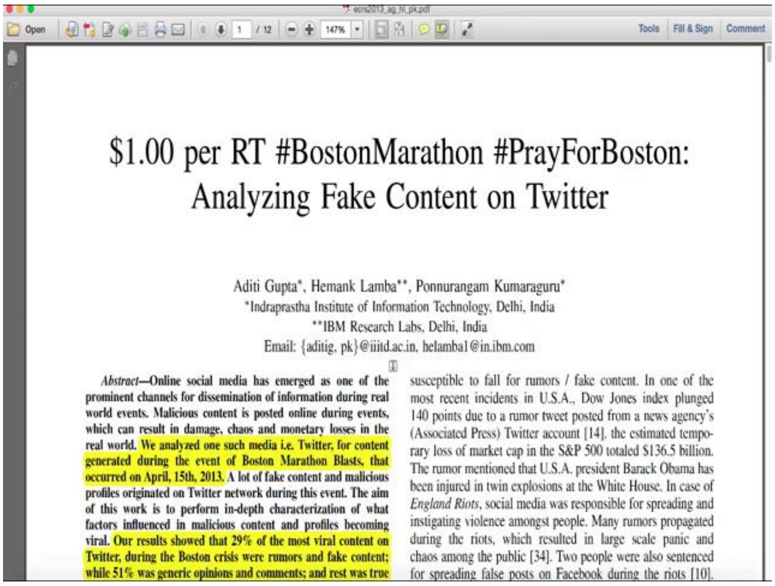
Privacy and Security in Online Social Media

Prof. Ponnurangam Kumaraguru ("PK")
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Week - 10.3
Lecture - 34
Boston Marathon Analyzing Fake Content on Twitter

Welcome back to the course on Privacy and Security in Online Social Media on NPTEL. This is week number 10, and we are going to look at continuing the trend.

(Refer Slide Time: 00:19)



$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

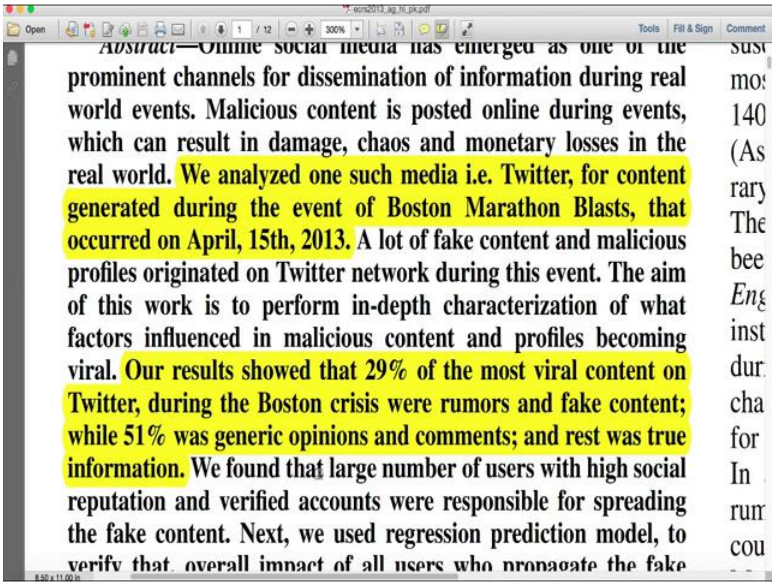
Abstract—Online social media has emerged as one of the prominent channels for dissemination of information during real world events. Malicious content is posted online during events, which can result in damage, chaos and monetary losses in the real world. We analyzed one such media i.e. Twitter, for 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

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].

We are going to look at another paper, which is also looking at some of the security and privacy issues on online social media. This paper is dollar one per RT Boston Marathon, Pray for Boston analyzing fake content on twitter. So, I briefly mentioned this in this section credibility, and trust, and if you remember this is one of the tweet that I used in example also to say please RTs this tweet, we will pay one dollar to Boston Marathon.

And you know all the otherwise the content in the title, hash tag is Boston Marathon, pray for Boston, I think in the later events also, you would have seen such kind of hash tags, pray for Paris, which was actually also trending when the Paris attack was happened, when the Paris attack happened, analysing fake content on twitter. So, I am going to go back to the content like the credibility that we saw, but then I just showed you some graph and I moved on. Now, we are going to actually see it in detail in terms of just paper writing. Paper is going to be our focus of this section.

(Refer Slide Time: 02:13)



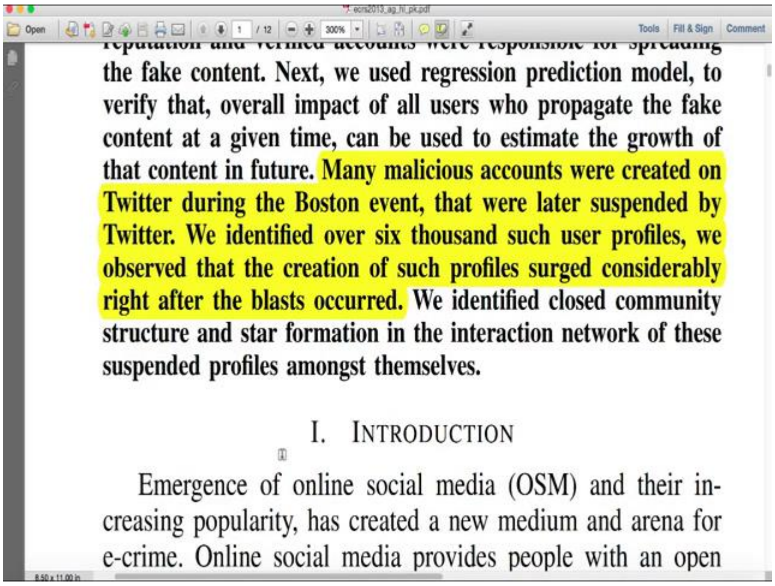
Abstract—Online social media has emerged as one of the prominent channels for dissemination of information during real world events. Malicious content is posted online during events, which can result in damage, chaos and monetary losses in the real world. We analyzed one such media i.e. Twitter, for 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

sus
mos
140
(As
rary
The
bee
Eng
inst
dur
cha
for
In
rum
cou

So, you should look at the claims by this paper. We analysed one such media, twitter please remember this was actually published in two thousand thirteen, two thousand twelve during that period. So, some of the mentions about the social media may be very very basic, we analysed one such media twitter for content generated during the event of Boston Marathon Blasts that occurred on April fifteenth two thousand thirteen.

So, what are the results that the authors claim is that twenty nine percent of the most viral content on twitter during the Boston Marathon crisis were rumours and fake content. So, this is how did they do it, they basically did took the largest, large set of rumours, large set of posts, asked users to annotate it and they also got the true positive rumours from another source, looked at how much of the total content generated about the Boston Marathon had these posts. And therefore, claim that around 29 percent of the most viral content were rumours. While 51 percent was generic opinions and comments and the rest was true information. So, this is the result of annotation that they did with the posts.

(Refer Slide Time: 03:59)



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

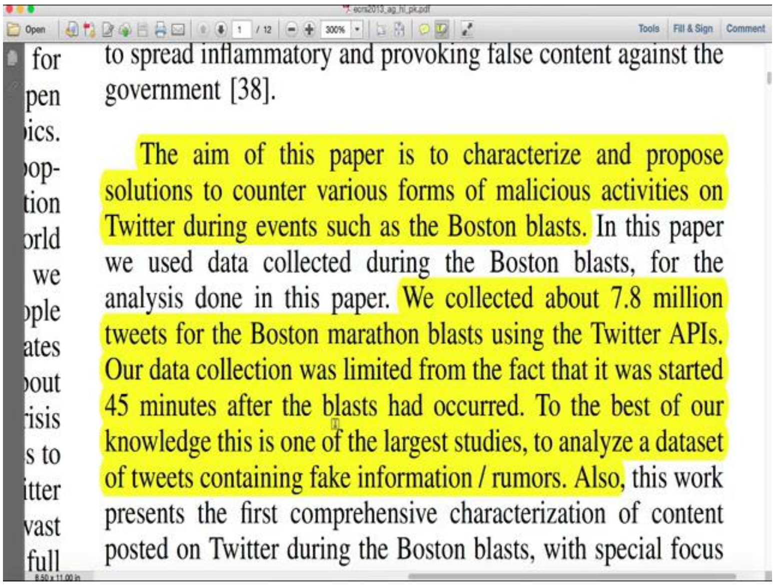
Emergence of online social media (OSM) and their increasing popularity, has created a new medium and arena for e-crime. Online social media provides people with an open

Many malicious accounts were created. So, there was a lot of malicious accounts that were created during the even just after the event during the Boston event that were later suspended by twitter. The authors identified over 6,000 such user profiles we observed that the creation of such profiles surged considerably right after the blast occurred. So, that is if you remember again one of the themes that I had been mentioning across the course is actually finding out these profiles.

How can you identify these profiles as fake profiles, which is one of main goals of services like twitter and it is not that easy also right given that this Boston Marathon blast has happened and people are under panic, the criminals are actually making use of this situation in terms of creating the account and posting content, which also, which are rumours and getting a lot of traction with it also.

As we have discussed before, if somebody uses hash tag Boston Marathon, hash tag based or Boston then, it is going to be available to people who are interested in that hash tag. And therefore, getting some rumours, getting some misinformation through this hash tag becomes much simpler. That is the abstract of the paper. Now let is look at actually the introduction, methodology, data set, contributions and the results that they have. So, in the introduction, the authors actually mentioned generally about the Boston Marathon blast, what happened and how the incidence is planned out of the Boston marathon.

(Refer Slide Time: 03:59)



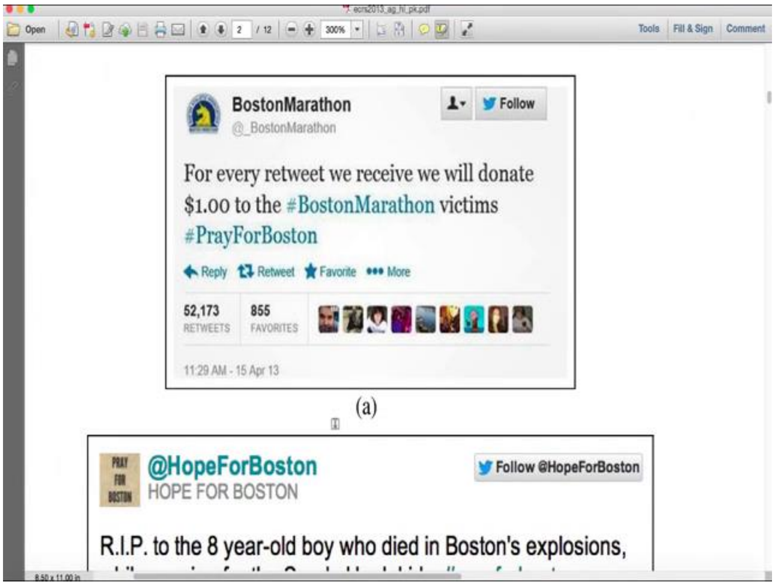
to spread inflammatory and provoking false content against the government [38].

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

So, 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 blast, right. So, the interest is to characterize, see and propose solution to counter various activities. So, one of the problems that they would take later is misinformation. Authors collected 7.8 million tweets; and later in methodology, we will see what kind of methodology, how did they actually collect 7.8 million posts. So, the data collection was limited from the fact that it was started 45 minutes after the blast. Interestingly, the way that these kinds of data collections are done is that an event happens and then, we see that there is an event, that has occurred and there is some hash tag with the events that is trending, take that hash tag, put it into the data collection and collect the data.

So, this probably has some draw back also, in this case it (Refer Time: 07:39). But the interesting thing is the 45 minutes thing again there, the author explained it later, but these 45 minutes is also interesting is because authors actually looked at the post that had come in these 45 minutes and some of them are actually retweets. So, the essentially the tweets that were tweeted even before 45 minutes is actually part of the data also here.

(Refer Slide Time: 08:22)



BostonMarathon
@_BostonMarathon

For every retweet we receive we will donate $1.00 to the #BostonMarathon victims #PrayForBoston

52,173 RETWEETS 855 FAVORITES

11:29 AM - 15 Apr 13

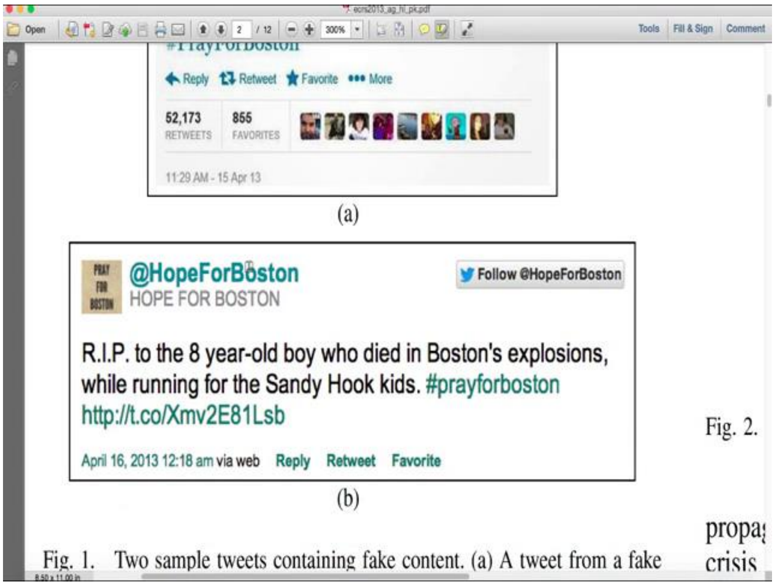
(a)

@HopeForBoston
HOPE FOR BOSTON

R.I.P. to the 8 year-old boy who died in Boston's explosions,

And one of the others claims by the paper is the largest dataset in terms of the rumours. So, here are the tweets that I have shown you in a different context before, which is for every retweet we receive ,we will donate 1 dollar to Boston Marathon victims, pray for Boston. So, the handle is underscore Boston Marathon, that is not as real account.

(Refer Slide Time: 08:38)



#PrayForBoston

52,173 RETWEETS 855 FAVORITES

11:29 AM - 15 Apr 13

(a)

@HopeForBoston
HOPE FOR BOSTON

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

April 16, 2013 12:18 am via web Reply Retweet Favorite

(b)

Fig. 2.

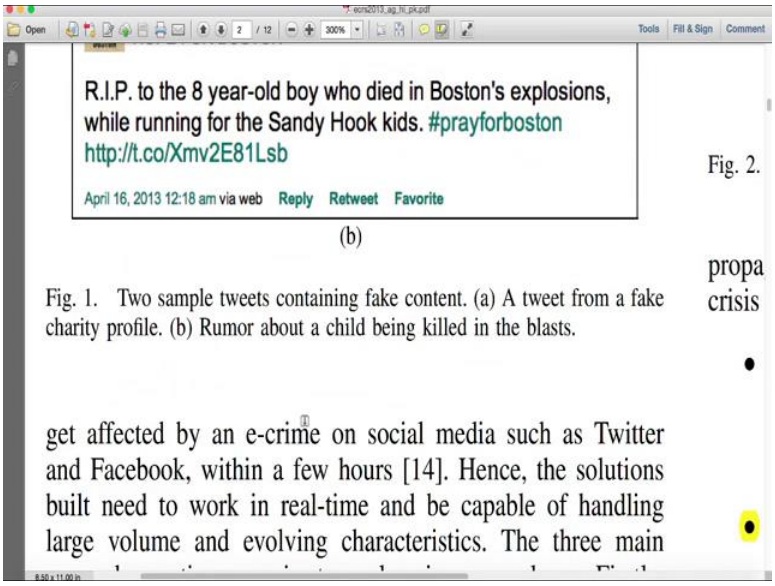
propa

crisis

Fig. 1. Two sample tweets containing fake content. (a) A tweet from a fake

Hope for Boston, rip to the 8 year old boy who died in Bostons explosions while running for the sandy hook kids to pray for Boston. Nothing, there was not a kid, a kid was not part of the Boston Marathon at all.

(Refer Slide Time: 09:06)



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

April 16, 2013 12:18 am via web Reply Retweet Favorite

(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

Fig. 2.

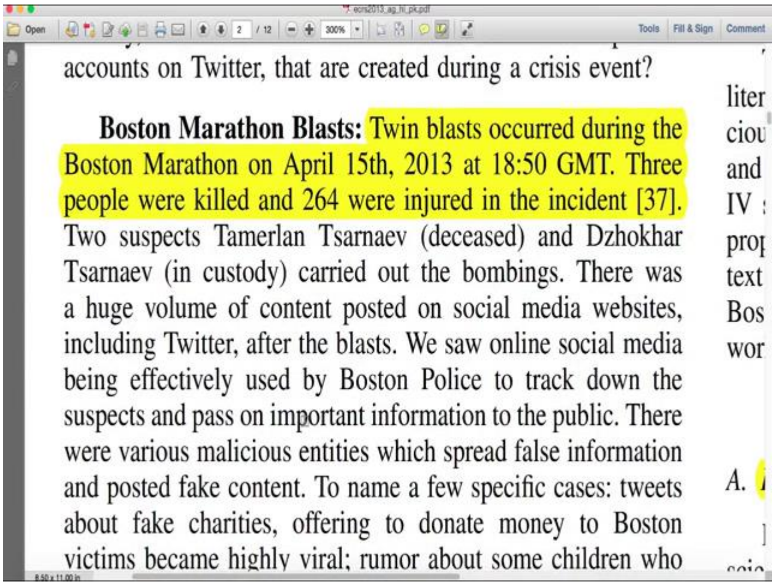
propa
crisis

•

•

There was no kid who took part in Boston Marathon and who actually was killed, because of the blast.

(Refer Slide Time: 09:17)



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

liter
ciou
and
IV
prop
text
Bos
wor

A.

So, this is just back ground of the Boston Marathon itself, it is a twin blast occurred during Boston Marathon, April 15th 2013 at about 18:50 GMT. 3 people were killed and 264 were injured in the incident.

(Refer Slide Time: 09:47)

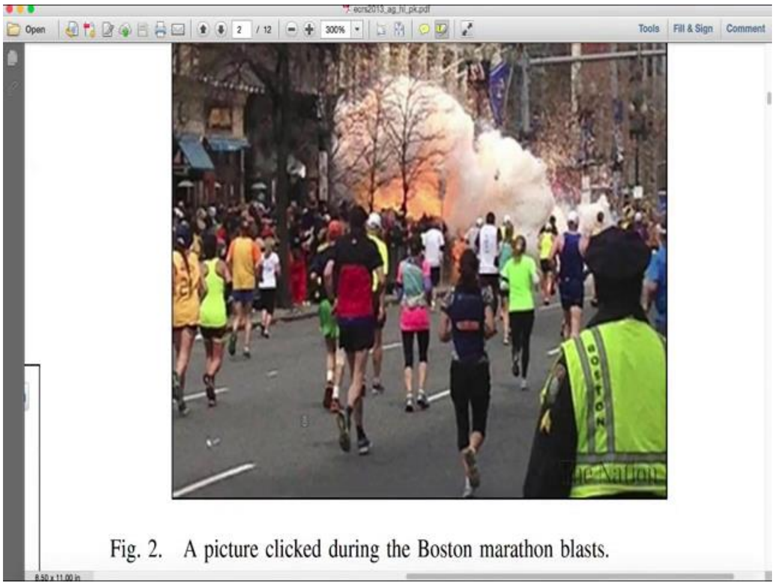
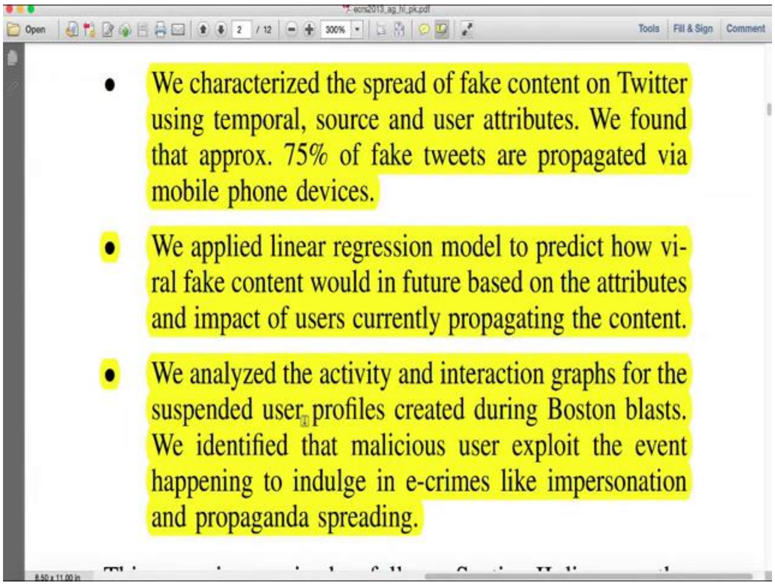


Fig. 2. A picture clicked during the Boston marathon blasts.

So, now looking at, so here is a picture that actually came out to the blast and this was one of the first pictures clicked during the Boston Marathon blast.

(Refer Slide Time: 09:52)

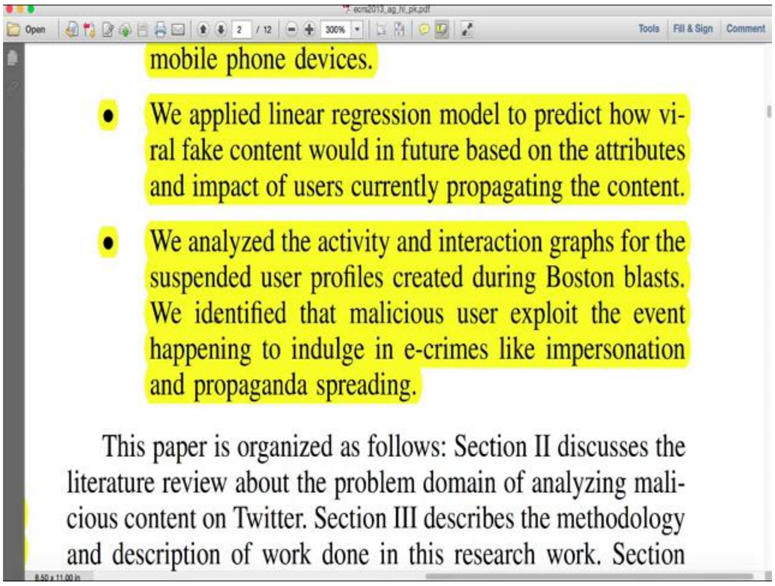


- We characterized the spread of fake content on Twitter using temporal, source and user attributes. We found that approx. 75% of fake tweets are propagated via mobile phone devices.
- We applied linear regression model to predict how viral fake content would in future based on the attributes and impact 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.

So, the authors actually characterized the spread of fake content on twitter using temporal source and user attributes. So, you will find some of the analysis in this paper pretty simple, which is the authors will just show you what kind of devices that the users had when they posted the content. And because again, keep the year in mind this is about two thousand thirteen. So, there is kind of the way that the authors look at it is slightly preliminary also.

Also we found that approximately 75 percent of the fake tweets are propagated via mobile phones. We applied linear regression model to predict how viral fake content in future based on attributes and impact of users is currently propagating the content. We analysed the activity and interaction graphs for the suspended user profiles, one interesting thing that they did was the authors actually looked at the suspended user profiles and did analysis of what kind of users at these used accounts were suspended.

(Refer Slide Time: 11:18)



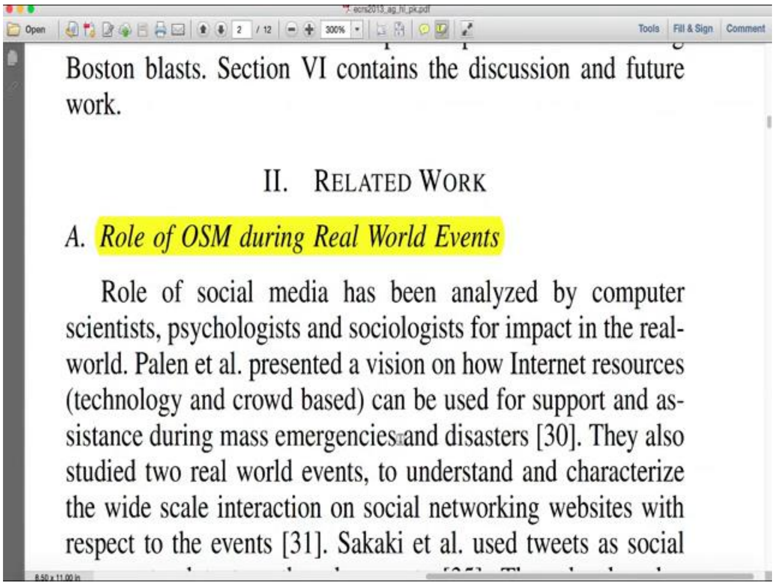
mobile phone devices.

- 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

I think the last line here I have said multiple times in this course, authors identified that malicious users exploit the event happening to indulge in e-crimes like impersonation and propaganda spreading.

(Refer Slide Time: 11:34)



Boston blasts. Section VI contains the discussion and future work.

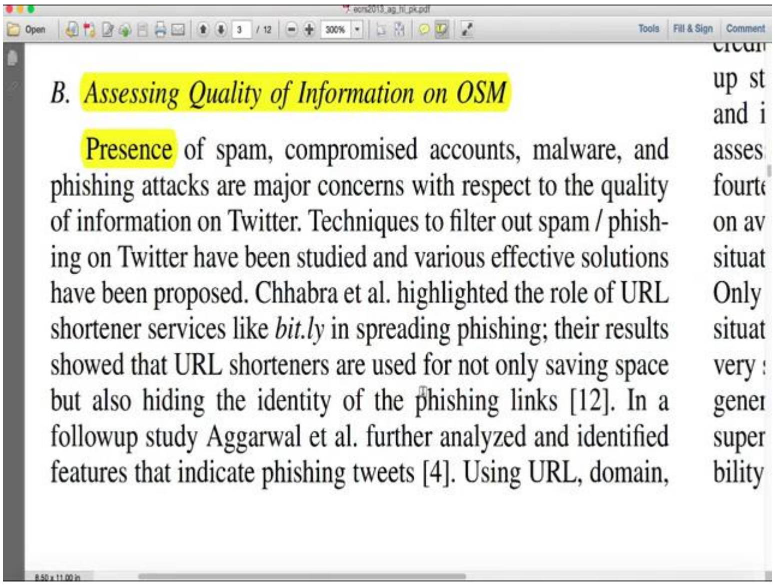
II. Related Work

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

Others broke the related work into multiple categories, the first one is the role of OSM, during real world events. I think it is very very clear that real world events, when it happens it actually triggers the online social media content generation, which is when there is earthquake. For example, there is Nepal earthquake, IPL cricket match all of this is actually triggered and huge amount of content that is generated on online social media.

(Refer Slide Time: 12:22)



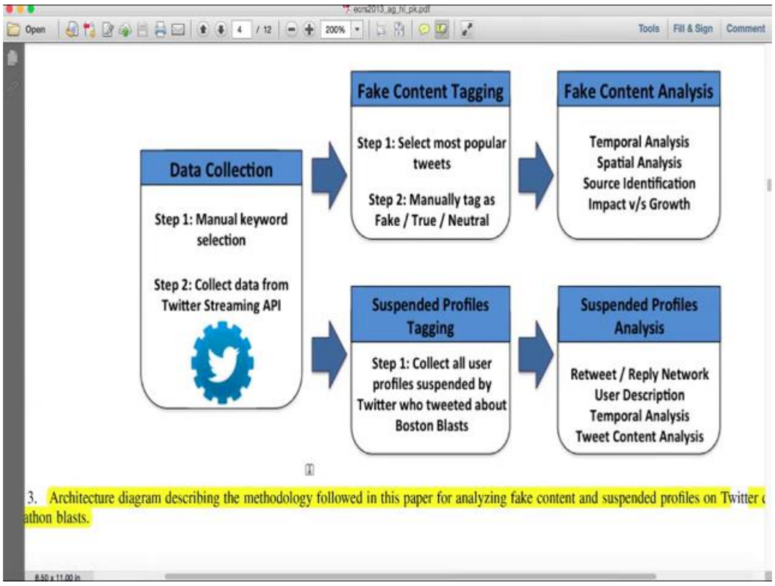
B. Assessing Quality of Information on OSM

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,

up st
and i
asses
fourt
on av
situat
Only
situat
very :
gener
super
bility

So, again the authors go through in detail about the different types of events where social media is being involved. And then, during the other category of events, the other category of literature that is available is actually assessing quality of information on online social media, which is to look at the misinformation, which is to look at what kind of content gets spread during this event, how characterization of them and how good or bad they are and topics around it. There are even books now, which actually look at this concept of misinformation on social media, big data. Next part gives you sense of what the related literature is, that you need to understand for this particular paper.

(Refer Slide Time: 13:27)



Data Collection

Step 1: Manual keyword selection

Step 2: Collect data from Twitter Streaming API

Fake Content Tagging

Step 1: Select most popular tweets

Step 2: Manually tag as Fake / True / Neutral

Fake Content Analysis

Temporal Analysis
Spatial Analysis
Source Identification
Impact v/s Growth

Suspended Profiles Tagging

Step 1: Collect all user profiles suspended by Twitter who tweeted about Boston Blasts

Suspended Profiles Analysis

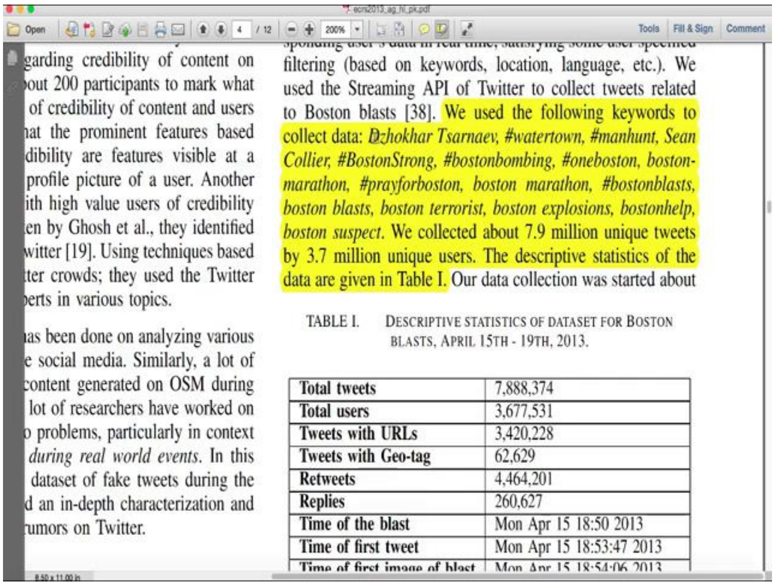
Retweet / Reply Network
User Description
Temporal Analysis
Tweet Content Analysis

3. Architecture diagram describing the methodology followed in this paper for analyzing fake content and suspended profiles on Twitter (
athon blasts.

Here is a very simple architecture that the authors had in terms of actually collecting and analyzing the data, which is keyword selection, manual keyword selection that is the forty five minutes delay that I said. And collect data from our twitter streaming API which all of you know, fake content tagging, select most popular tweets, manually tag as fake, true or neutral. So, some kind of annotations happens with the tweets that is from the event, suspended profiles tagging, collect all user profiles suspended by twitter who tweeted about Boston Marathon.

So, all the tweets that were, all the users that were suspended who had Boston Marathon hash tag were snipped out. Fake and temporal, spatial, source, impact and suspended profile analysis, which is retweet reply network, user description, temporal analysis. Essentially the two parts of analysis is done. One is the content from the event to understand the fake or misinformation, the other one is the suspended profile users.

(Refer Slide Time: 14:53)



garding credibility of content on out 200 participants to mark what of credibility of content and users hat the prominent features based dibility are features visible at a profile picture of a user. Another ith high value users of credibility en by Ghosh et al., they identified witter [19]. Using techniques based tter crowds; they used the Twitter perts in various topics.

has been done on analyzing various e social media. Similarly, a lot of content generated on OSM during lot of researchers have worked on o problems, particularly in context *during real world events*. In this dataset of fake tweets during the d an in-depth characterization and rumors on Twitter.

sponding 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, boston-marathon, #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

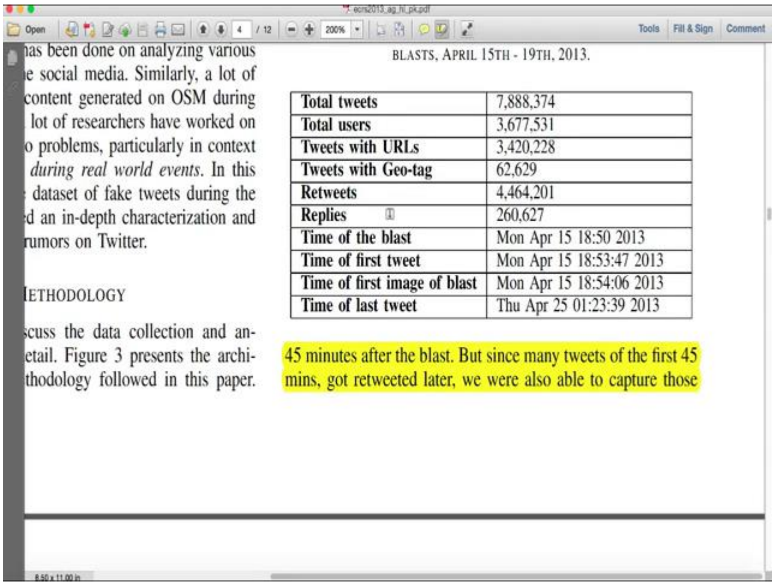
TABLE I. DESCRIPTIVE STATISTICS OF DATASET FOR BOSTON BLASTS, APRIL 15TH - 19TH, 2013.

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

So, if you look at the data collection, so this is what we said earlier, how do they collect this 7.9 million, they used the keywords called the person who was actually suspected, hash tag watertown, hash tag manhunt, Sean Collier, hash tag BostonStrong hash tag bostonbombing, hash tag oneboston, boston hyphen marathon, hash tag prayforboston, boston marathon with the space, hash tag boston blasts, boston blast with the space, boston terrorist with the space, boston explosions with the space, bostonhelp without a space, boston suspect.

So, using all this is there is an interesting way of actually collecting all this. If you just know that hash tag boston marathon is what you are looking for, you take that hash tag collects some hundred tweets, look at the frequency of the count of other words and then, take that and then put them back into the search query and start collecting the data for that, so that is the kind of I mean in information retrieval kind of a topics, this is referred as query expansion. You can take one query and then expand it as you would like.

(Refer Slide Time: 16:31)



has been done on analyzing various
e social media. Similarly, a lot of
content generated on OSM during
lot of researchers have worked on
o problems, particularly in context
during real world events. In this
dataset of fake tweets during the
d an in-depth characterization and
rumors on Twitter.

IETHODOLOGY

scuss the data collection and an-
etail. Figure 3 presents the archi-
thodology followed in this paper.

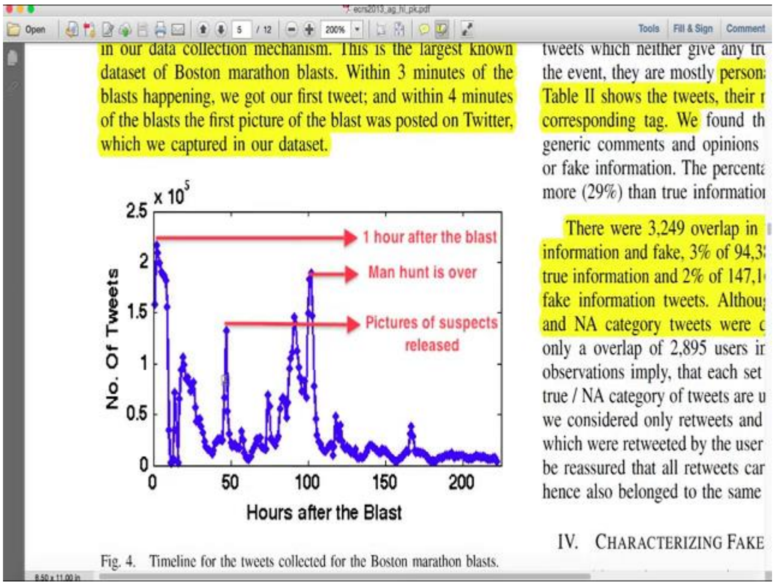
BLASTS, APRIL 15TH - 19TH, 2013.

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

This data is about 7.9 million, and total tweets, total users, total URLs, tweets with geo tag, retweets, replies, time of the blast, time of the first tweet, time of the first image of blast, time of last tweet. Basically, to just show you what distribution that the data has, the 45 minutes blast so as I explained before, since in our data set collection, this is the largest known data set. So, they have explained it here, but since many tweets of the 45 minutes got retweeted later.

(Refer Slide Time: 17:10)



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.

tweets which neither give any tru
the event, they are mostly person
Table II shows the tweets, their r
corresponding tag. We found th
generic comments and opinions
or fake information. The percenta
more (29%) than true informatio

There were 3,249 overlap in
information and fake, 3% of 94,3
true information and 2% of 147,1
fake information tweets. Althou
and NA category tweets were c
only a overlap of 2,895 users in
observations imply, that each set
true / NA category of tweets are u
we considered only retweets and
which were retweeted by the user
be reassured that all retweets car
hence also belonged to the same

IV. CHARACTERIZING FAKE

We were also able to capture to those in our data collection mechanism. This is the largest known dataset of Boston marathon blast. More than 3 minutes of the blast happening, we got our first tweet. So, there is also this interesting phenomenon where you will actually see that the real world, that the content or the frequency of appearance of social media content is actually directly related to the events that are happening in real-world; in real-world and the social media frequency are actually very much correlated.

So, here if you see, x-axis is the hours after the blast, y-axis is the number of tweets, and wherever the peak has happened, there is a direct relation to the real world event. The first one being 1 hour after the blast, the second one being pictures of suspects released and the third one being Man hunt is over.

(Refer Slide Time: 18:17)

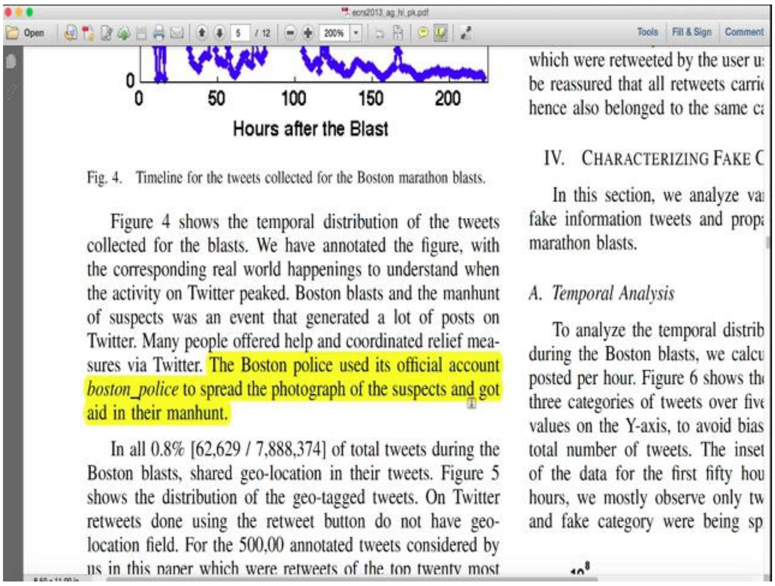


Fig. 4. Timeline for the tweets collected for the Boston marathon blasts.

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

which were retweeted by the user u
be reassured that all retweets carri
hence also belonged to the same c

IV. CHARACTERIZING FAKE C

In this section, we analyze va
fake information tweets and prop
marathon blasts.

A. Temporal Analysis

To analyze the temporal distrib
during the Boston blasts, we calcu
posted per hour. Figure 6 shows th
three categories of tweets over fiv
values on the Y-axis, to avoid bias
total number of tweets. The inset
of the data for the first fifty hou
hours, we mostly observe only tw
and fake category were being sp

The Boston police used its official account to spread the photograph of the suspects and got aid in their manhunt. So, one of the things that we actually analyzed, the authors actually analyzed during the event was that Boston marathon, the legitimate account started sharing pictures through the legitimate account to get inputs from the public itself which was an interesting behaviour by the police organisation.

(Refer Slide Time: 18:47)

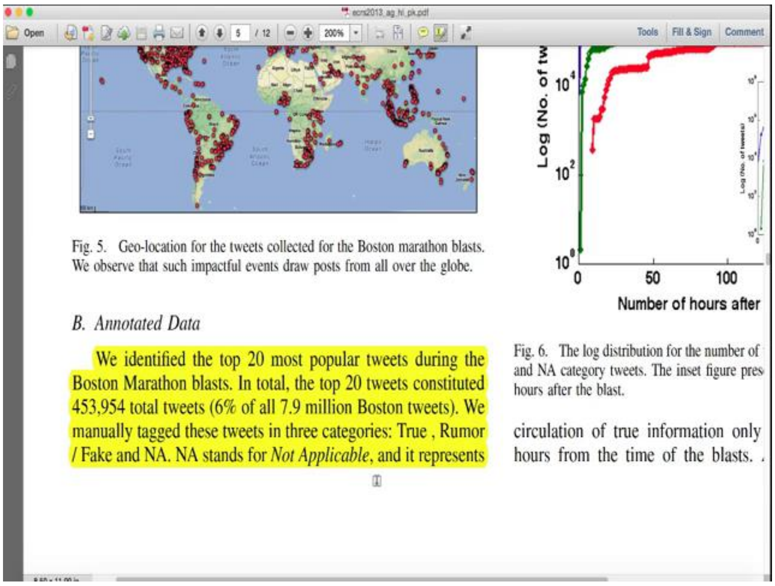


Fig. 5. Geo-location for the tweets collected for the Boston marathon blasts. We observe that such impactful events draw posts from all over the globe.

B. Annotated Data

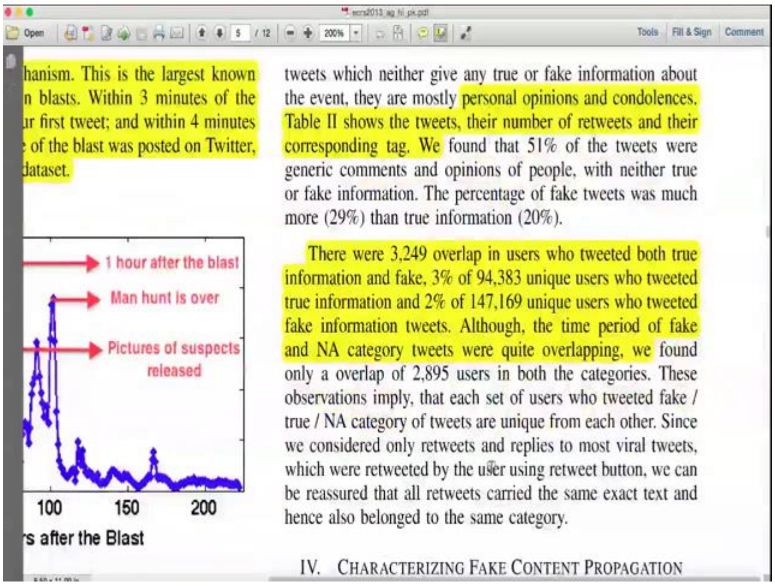
We identified the top 20 most popular tweets during the Boston Marathon blasts. In total, the top 20 tweets constituted 453,954 total tweets (6% of all 7.9 million Boston tweets). We manually tagged these tweets in three categories: True , Rumor / Fake and NA. NA stands for *Not Applicable*, and it represents

Fig. 6. The log distribution for the number of and NA category tweets. The inset figure pres hours after the blast.

circulation of true information only hours from the time of the blasts.

So, the annotations that the authors did was they identified the top 20 most popular tweets during the Boston Marathon. In total, the 20 tweets constituted of 400,000 total tweets which is 6 percent. Authors manually tagged these tweets in three categories true, rumour, fake and not available, NA stands for not applicable.

(Refer Slide Time: 19:27)



hanism. This is the largest known n blasts. Within 3 minutes of the ur first tweet; and within 4 minutes of the blast was posted on Twitter, dataset.

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%).

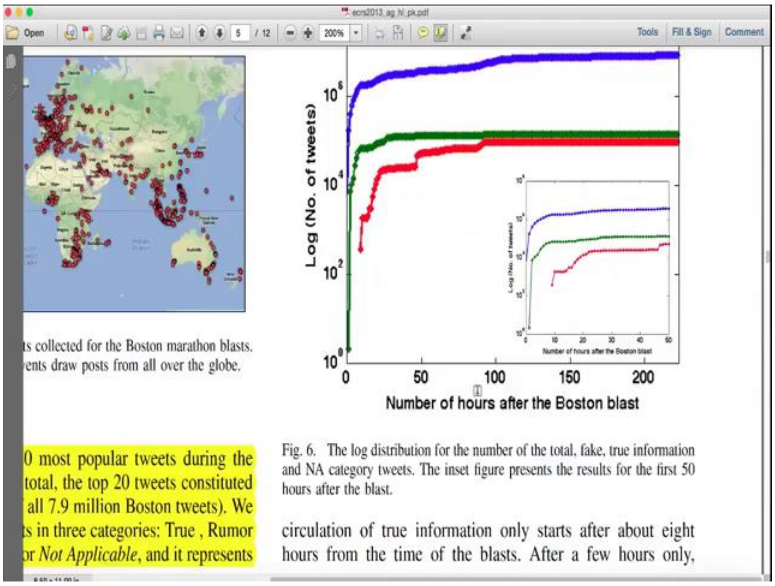
There were 3,249 overlap in users who tweeted both true information and fake, 3% of 94,383 unique users who tweeted true information and 2% of 147,169 unique users who tweeted fake information tweets. Although, the time period of fake and NA category tweets were quite overlapping, we found only a overlap 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

And it represents tweets which neither gives any true information or fake information about the event. They are mostly personal opinions and condolences. So, you should remember in the abstract we saw that about 50 percent of them are actually of this category. So, 51 of this percent of these were general comments and opinions. And the percentage of fake tweets was much more, 29 percent than true information about 20 percent, so that is the outcome of the annotation. So, if you remember, many of the papers that we have seen, many of research that we have seen, there is this human annotation done to learn the model to understand, how the real users react to these posts.

There were about 3,249 overlap in the users who tweeted true information and fake, 3 percent of 94000 unique users who tweeted true information, and 2 percent of 147,000 unique users who tweeted fake information tweets. So, essentially this is saying that there are people who are posting both real and fake information, there are people who are 2 percent of the total number of unique users who posted fake information also.

(Refer Slide Time: 20:58)



ts collected for the Boston marathon blasts. ents draw posts from all over the globe.

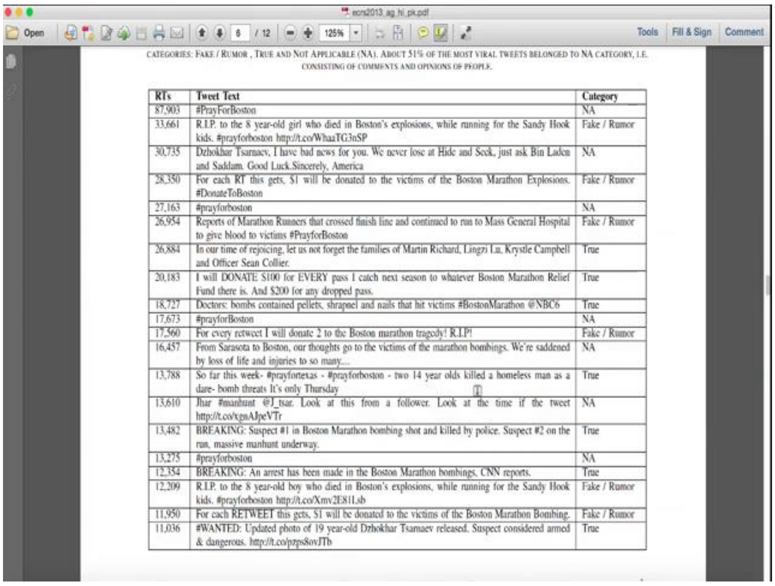
0 most popular tweets during the total, the top 20 tweets constituted all 7.9 million Boston tweets). We ts in three categories: True , Rumor or *Not Applicable*, and it represents

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,

Temporal analysis, this is the graph that I have shown you in one of the very earlier lectures I think about week 2 or something, where I kind of described what the spread of the rumour is. If we look at this graph, I will make it short here true information is actually starting much later than the rumour information which is green colour and rumour is also studying much faster than the real information. So, there are ways to actually reduce this, one is to bring this real information earlier it starts early and then and the other solution is to get the rumour information fall flat very soon.

(Refer Slide Time: 21:59)

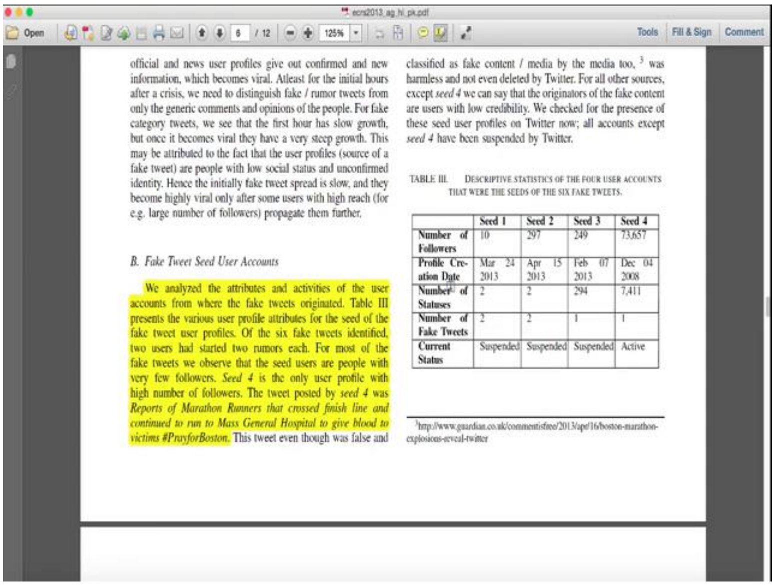


CATEGORIES: FAKE / RUMOR , TRUE AND NOT APPLICABLE (NA). ABOUT 51% OF THE MOST VIRAL TWEETS BELONGED TO NA CATEGORY, I.E. CONSISTING OF COMMENTS AND OPINIONS OF PEOPLE.

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

If these techniques can be achieved then I think the spread of a misinformation can be reduced a lot. Here is a table which actually gives you the full details about the events that authors actually collected, posts that they had which is tweet text, RTs and category. So this is the outcome of the manual annotation, and the authors are just giving you a sample of all the RTs that they had which is in the decreasing order here, the category here. So, the top RT was hash tag PrayForBoston which had about 87000, and it is actually, nothing is available right PrayForBoston, you cannot really say whether it is actually malicious, legitimate or it has no information.

(Refer Slide Time: 22:54)



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

So, now what authors did was they actually looked at, if you remember again, one of the contribution in the paper is to do analyse the user accounts that were actually suspended. So, what they do is they analyse the attributes and activities of the user accounts from where the fake tweets originated. Table 3 presents the various user profiles attributes from the seed of the fake tweet users, user profiles of the 6 fake tweets identified, 2 users had started two rumours each, for most of the fake tweets we observe that the seed users are people with few followers. Seed 4 is the only user profile with the high number of followers. The tweet posted by seed 4 was reports of Boston marathon runners that crossed finish line and continued to run to Mass General Hospital to give blood to victims.

So, here are for accounts that this paragraph was talking about. The numbers of followers, if you see the first three seeds that are pretty small. So, the point is that the rumours could be created by follow accounts, who had followers very small. But it gets spread fastly because somebody in the chain, in the food chain of looking at these tweets, who got more popular also looks at the tweets and post it, so that is ne of the problem that you can actually look at.

Also how to actually reduce this particular post, from this particular rumour, from a user who is not, who does not have a large number of followers, going into the hands of people who are actually popular and actually retweeting that. But seed 4 if you see, is actually 73000 followers these are the only accounts which are actually more popular and it is also active as per the data collection that was done.

(Refer Slide Time: 25:05)

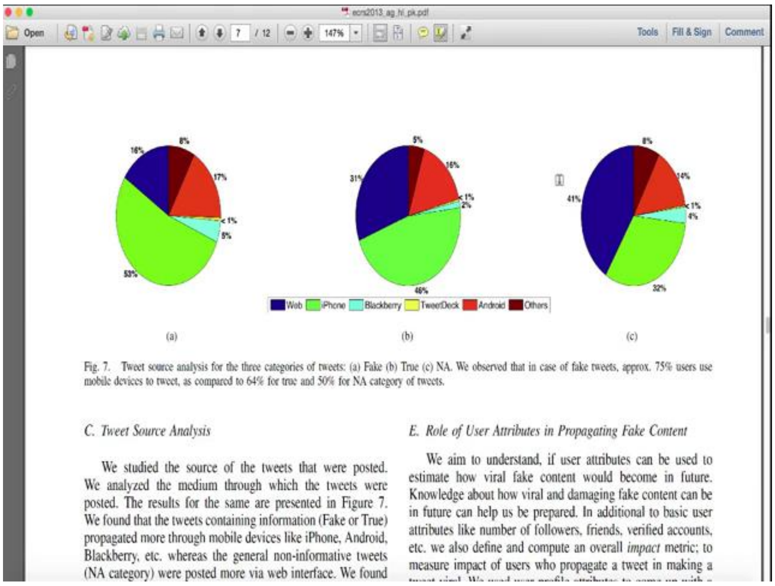


Fig. 7. Tweet source analysis for the three categories of tweets: (a) Fake (b) True (c) NA. We observed that in case of fake tweets, approx. 75% users use mobile devices to tweet, as compared to 64% for true and 50% for NA category of tweets.

C. Tweet Source Analysis

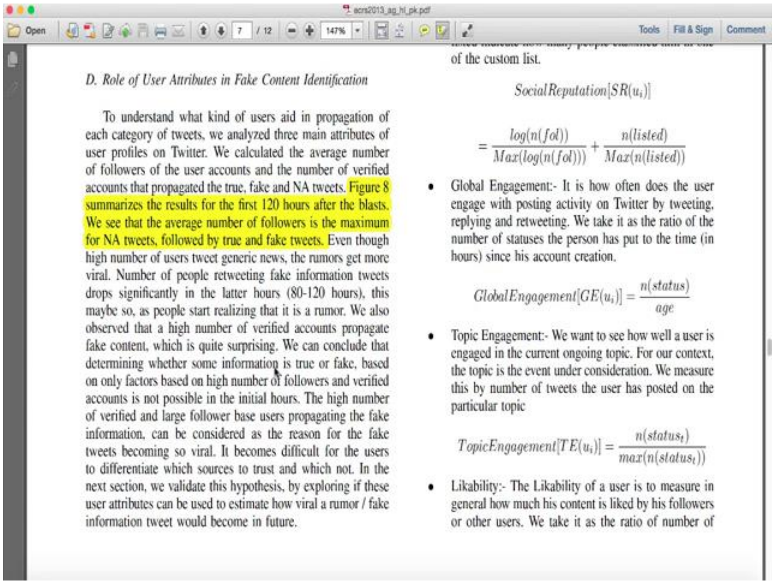
We studied the source of the tweets that were posted. We analyzed the medium through which the tweets were posted. The results for the same are presented in Figure 7. We found that the tweets containing information (Fake or True) propagated more through mobile devices like iPhone, Android, Blackberry, etc. whereas the general non-informative tweets (NA category) were posted more via web interface. We found

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

Here is the graph that I said before which is that authors actually looked at the true malicious and other information that is available through that device, that which device was used. So, a - is all the fake information, in all the fake information it looks like majority of them are actually using iphone. And true information also it seems to be the same whereas the web is actually used to more in terms of the true information versus the fake information. And if you look at no information that is available, the web seems to be the highest. Well this is generally only to get a sense of what kind of device people are using while posting.

(Refer Slide Time: 26:09)



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.

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

And authors also, tweet source analysis which is what I said. Now let us look at the analysis of role of user attributes in fake content identification, which is what are the features in the user in terms of followers, in terms of whether the account is verified, how do they relate to the fake content that is getting propagated. The figure 8 actually shows you the content on the x-axis till 0 to 120 hours, y-axis to be the number of followers. It will clearly show that the first one is actually at the highest one, is at the not applicable tweet, then followed by the true content, and then followed by the fake contents, which is number of followers, the total number of followers or the followers for the tweets that got propagated, which is take at the least number of followers.

(Refer Slide Time: 27:03)

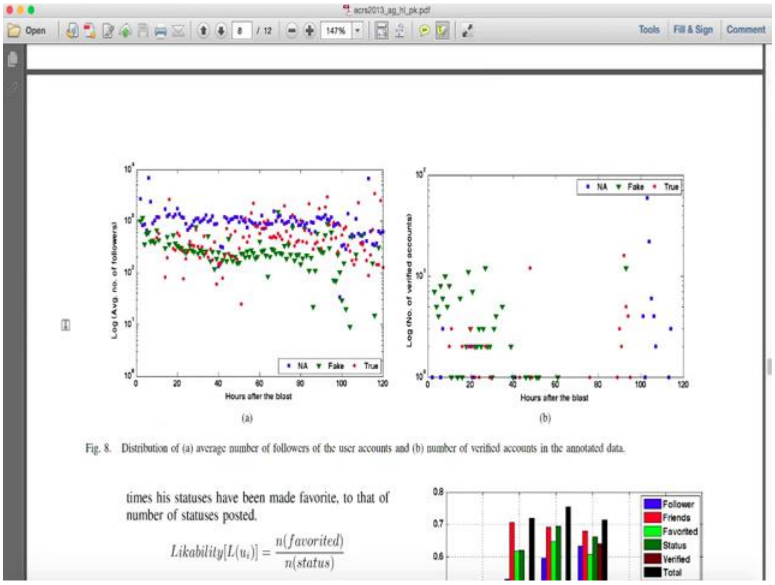


Fig. 8. Distribution of (a) average number of followers of the user accounts and (b) number of verified accounts in the annotated data.

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

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

That is what you will see here, x-axis to be the hours, y-axis to be the number of followers. And the green triangle is actually at the lowest showing that the followers to the fake content are the low. Then the red one, red circles which had the followers to the true content, and then the blue one which is actually the followers for the not applicable content. You should look the dots in the right, it actually shows you the verified accounts, verified accounts also show hours after the blast, there is a lot of content posted by verified accounts which are not applicable and true later.

Whereas, initially there are some content that are getting propagated by fake content, fake content is getting propagated by people who are actually verified accounts. So, there is a problem as I said where initially people do not verify the content, people do not check what is going on, they have a verified account, they are just pushing this content and therefore, this content gets propagated much faster.

(Refer Slide Time: 28:30)

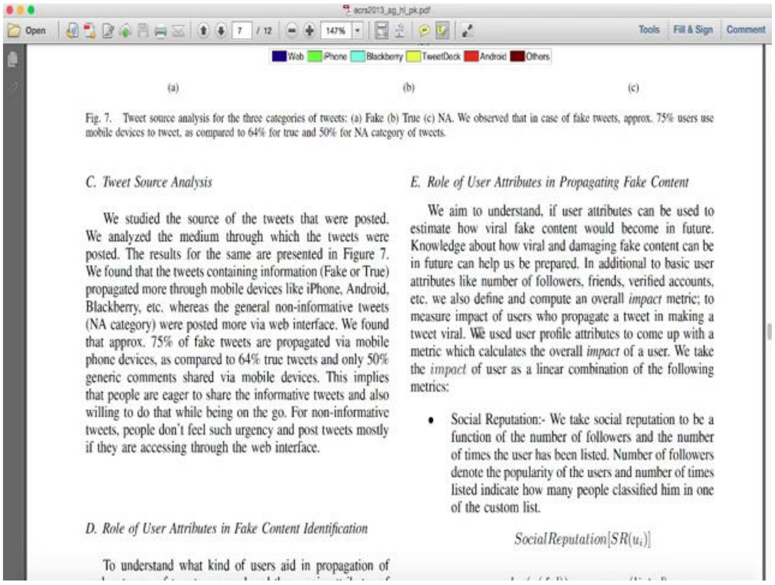


Fig. 7. Tweet source analysis for the three categories of tweets: (a) Fake (b) True (c) NA. We observed that in case of fake tweets, approx. 75% users use mobile devices to tweet, as compared to 64% for true and 50% for NA category of tweets.

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

E. Role of User Attributes in Propagating Fake Content

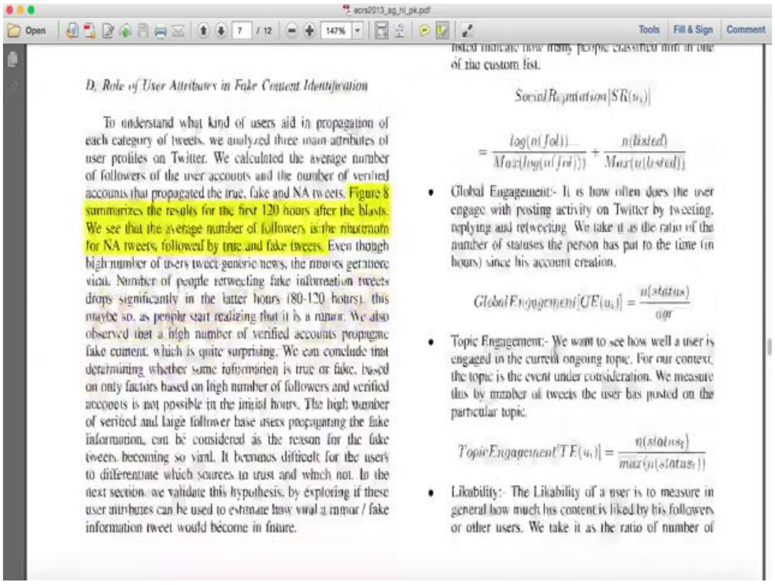
We aim to understand, if user attributes can be used to estimate how viral fake content would become in future. Knowledge about how viral and damaging fake content can be in future can help us be prepared. In additional to basic user attributes like number of followers, friends, verified accounts, etc. 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.

$$SocialReputation[SR(u_i)]$$

So, that is the role of user attributes in fake content identification and there is this whole set of analysis that one can do in terms of actually propagating, analysing the fake content.

(Refer Slide Time: 28:53)



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.

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

So, if you see here, there are concepts that are mentioned here. I will briefly tell you what they are, social reputation which is like the cloud global engagement, interactions with the other users, topic engagement, linkability, credibility, all of this, the authors basically what they did was they created, they had all these metrics SR, GE, TE, L and C which is in the reverse order credibility, linkability, topic engagement, global engagement and social reputation.

(Refer Slide Time: 29:21)

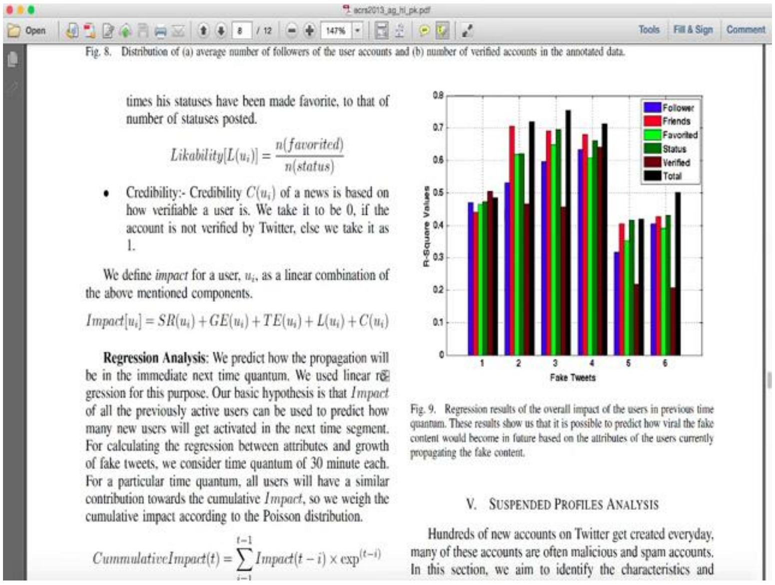


Fig. 8. Distribution of (a) average number of followers of the user accounts and (b) number of verified accounts in the annotated data.

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.

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

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.

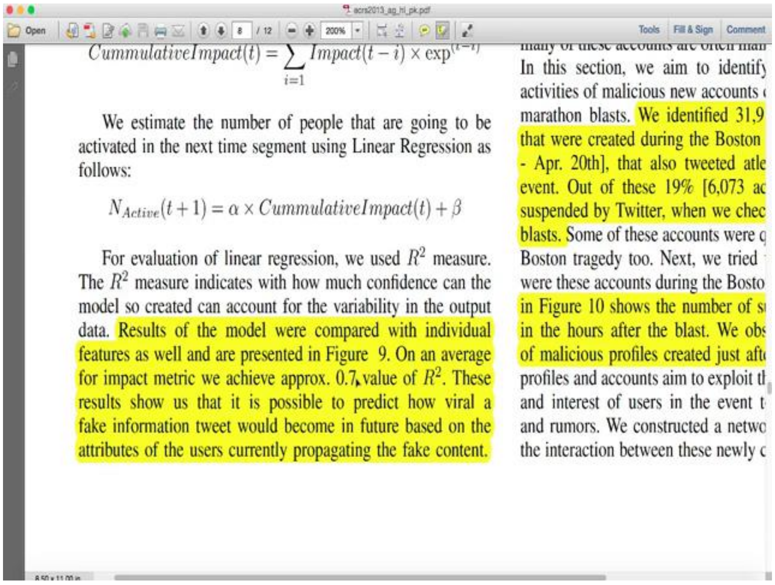
V. SUSPENDED PROFILES ANALYSIS

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

They used this and created a metric called impact and what they were trying to figure out is that whether we can use these features, the five, ones that I just now said to predict what impact of the post is going to be or virality of the post is going to be and things like that. The graph 9, the figure 9 on the right hand side here actually shows you that it is possible to find out the propagation of the content using all these features right.

So, what does this show, regression shows results of the overall impact to 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. So, they basically used followers, friends, favourite, status, verified, all of them find out whether these information can be used to find the virility of the fake content that is posted on twitter.

(Refer Slide Time: 30:45)



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

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

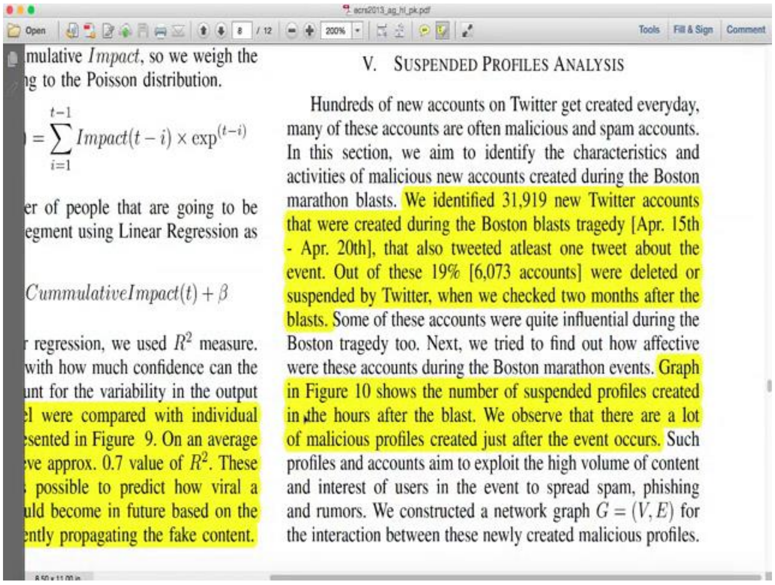
$$N_{Active}(t+1) = \alpha \times CummulativeImpact(t) + \beta$$

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.

many of these accounts are often mali
In this section, we aim to identify
activities of malicious new accounts c
marathon blasts. We identified 31,9
that were created during the Boston
- Apr. 20th], that also tweeted atle
event. Out of these 19% [6,073 ac
suspended by Twitter, when we chec
blasts. Some of these accounts were q
Boston tragedy too. Next, we tried
were these accounts during the Bosto
in Figure 10 shows the number of s
in the hours after the blast. We obs
of malicious profiles created just aft
profiles and accounts aim to exploit th
and interest of users in the event t
and rumors. We constructed a netwo
the interaction between these newly c

So, that is what they kind of showed that results of the model were compared with the individual features as well and are presented in figure 9. On an average for the impact metric, we achieve approximately 0.7 value of R square. These result show us that it is possible to predict how viral a fake information tweet would become in future based on the content attributes of the users currently propagating the fake content, so that is about the analysis.

(Refer Slide Time: 31:16)



mulative *Impact*, so we weigh the
ng to the Poisson distribution.

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

er of people that are going to be
egment using Linear Regression as

$$CummulativeImpact(t) + \beta$$

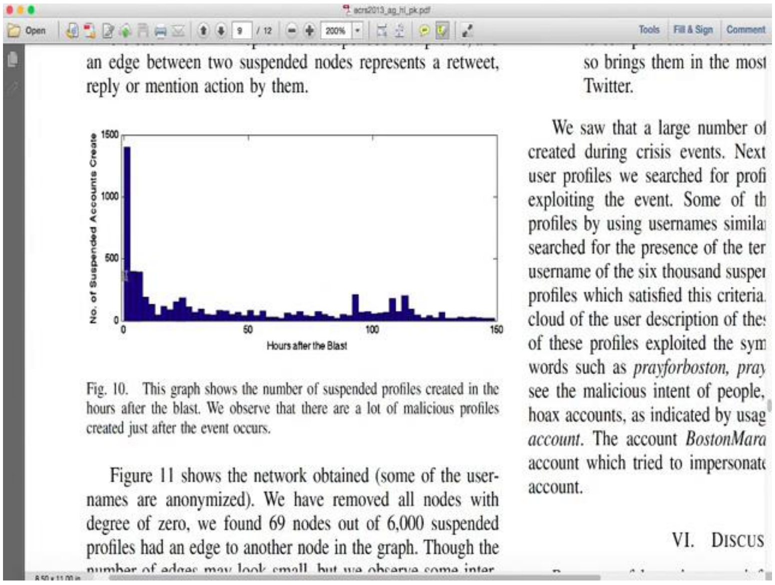
r regression, we used R^2 measure.
with how much confidence can the
unt for the variability in the output
el were compared with individual
esented in Figure 9. On an average
ve approx. 0.7 value of R^2. These
possible to predict how viral a
uld become in future based on the
ently propagating the fake 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.

So, now, let us look at the last analysis that the authors did was suspended profile analysis. They identified close to 32000 twitter accounts that were created during the blast. And out of these, 19 percent were deleted or suspended by twitter when we checked 2 months after the blast. So, basically since we had the post, post has a user, user has a id, we went back and checked, authors went back and checked the whether this particular user id is there, if the user id is not there, it was tagged that it was actually deleted.

(Refer Slide Time: 32:12)



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 inter

so brings them in the most Twitter.

We saw that a large number of created during crisis events. Next user profiles we searched for profi exploiting the event. Some of th profiles by using usernames similar searched for the presence of the ter username of the six thousand susper profiles which satisfied this criteria. cloud of the user description of thes of these profiles exploited the sym words such as *prayforboston, pray* see the malicious intent of people, hoax accounts, as indicated by usag *account*. The account *BostonMara* account which tried to impersonate account.

VI. DISCUS

So, figure 10 shows the number of suspended profile stated in hours after the blast. We observed that there are lot of malicious profiles created just after the event. If you look at this figure 10, this graph shows the number of suspended profiles. So, there were about 1400 profiles that were suspended which were just created few hours after the blast, right.

(Refer Slide Time: 32:28)

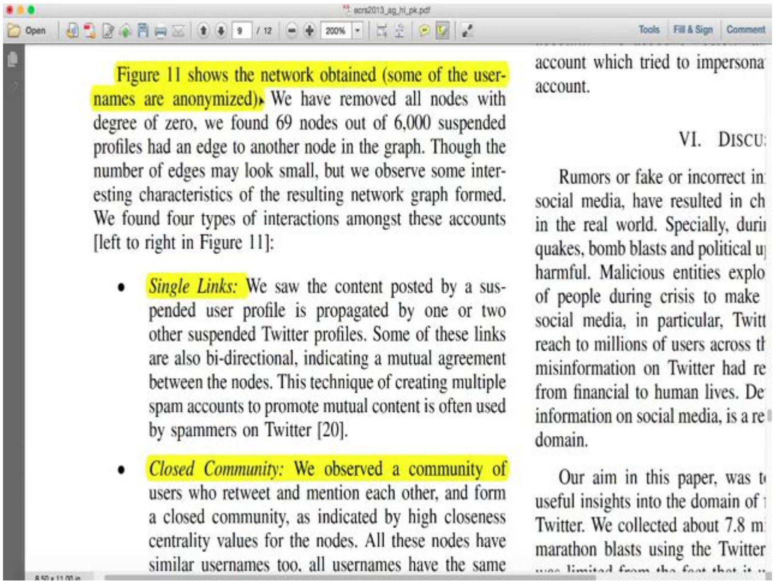


Figure 11 shows the network obtained (some of the usernames are anonymized). We have removed all nodes with degree of zero, we found 69 nodes out of 6,000 suspended profiles had an edge to another node in the graph. Though the number of edges may look small, but we observe some interesting characteristics 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

account which tried to impersona
account.

VI. Discu

Rumors or fake or incorrect in
social media, have resulted in ch
in the real world. Specially, duri
quakes, bomb blasts and political u
harmful. Malicious entities explo
of people during crisis to make
social media, in particular, Twitt
reach to millions of users across t
misinformation on Twitter had re
from financial to human lives. De
information on social media, is a re
domain.

Our aim in this paper, was t
useful insights into the domain of
Twitter. We collected about 7.8 m
marathon blasts using the Twitter
was limited from the fact that it w

And similarly you can also draw links between these handles. Figure 9 shows the networks obtained, some of the users' names are anonymized. They have removed all the nodes with the degree zero, we found that 69 nodes out of 6,000 suspended profiles had an edge to another node in the graph. This basically shows that the handles which are propagating fake content are actually collected among themselves also. Single links, so we found four types of interactions among these accounts, single links they are not interacting with anybody, closed community there is a closed community, so, here are the graphs.

(Refer Slide Time: 33:18)

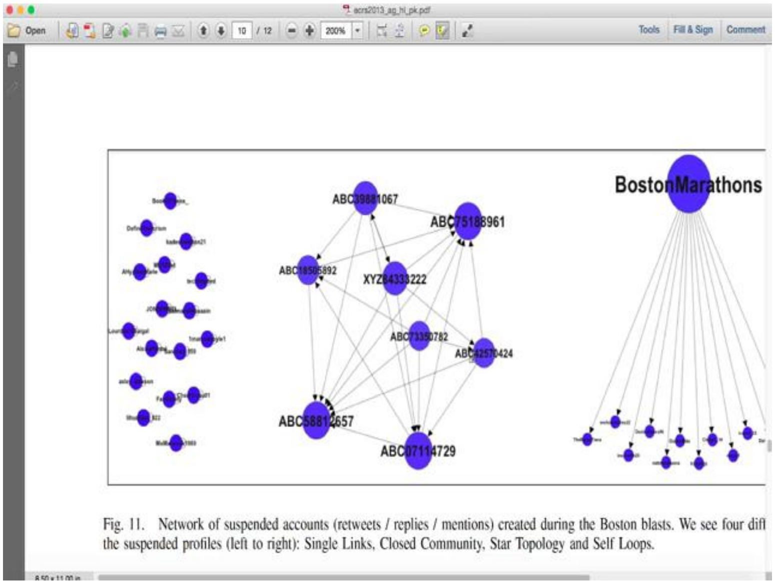
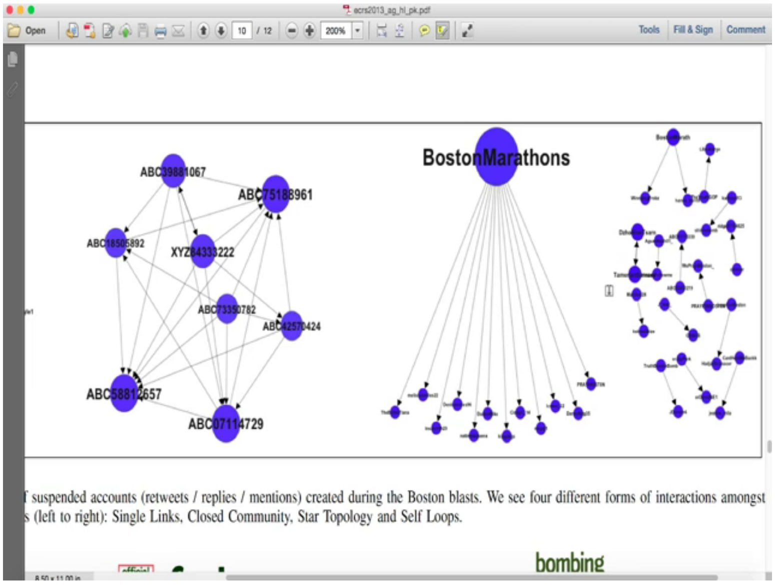


Fig. 11. Network of suspended accounts (retweets / replies / mentions) created during the Boston blasts. We see four diff
the suspended profiles (left to right): Single Links, Closed Community, Star Topology and Self Loops.

So, these are all singletons, and then there is this closed network of people, everybody is connected to everybody.

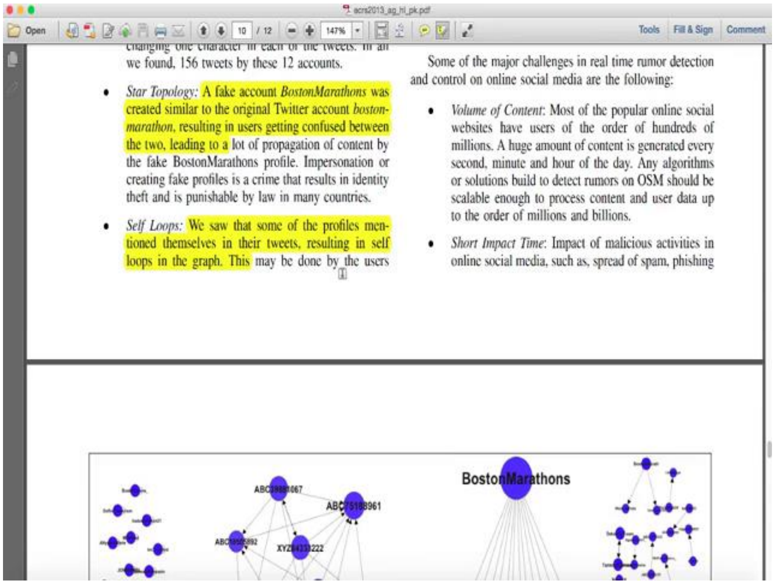
(Refer Slide Time: 33:27)



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

And the other one is, there is information spread only from one account to others and there is also this self loops which is, each of them spreading. So, I am connected to you, you are connected to me, and I am posting your content you are posting the content, right. So, that is the technically, there are single links on the left one, closed community star topology is the structure here and self loops are the ones here.

(Refer Slide Time: 34:06)



changing one character in each of the tweets. In all we found, 156 tweets by these 12 accounts.

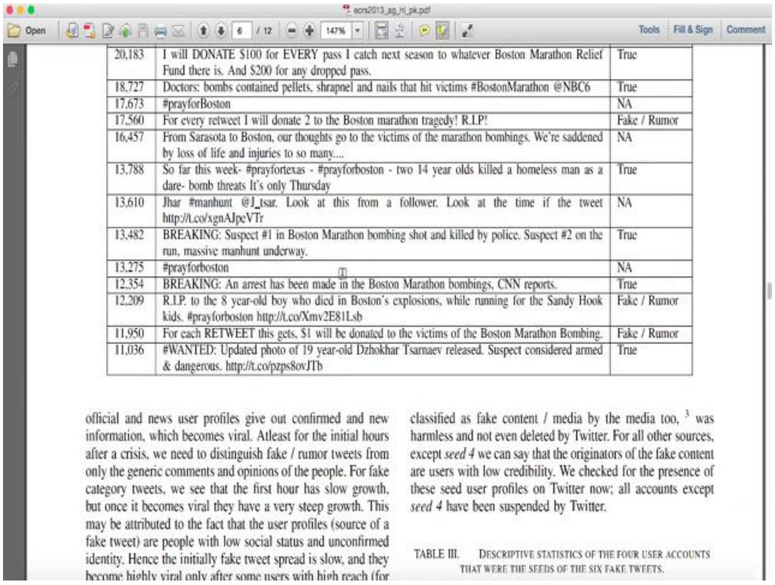
- *Star Topology:* A fake account *BostonMarathons* was created similar to the original Twitter account *boston-marathon*, resulting in users getting confused between the two, leading to a lot of propagation of content by the fake BostonMarathons profile. Impersonation or creating fake profiles is a 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

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

Self loops, there are some of the profiles mentioned themselves in their tweets resulting in self loops in the graphs.

(Refer Slide Time: 34:39)



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

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.

So, that is about the analysis that was done in this paper in terms of actually analysing the fake content propagation in twitter during the event Boston marathon. And this is just not comprehensive view of everything, it is just only one type of analysis, there have been many other papers which are looking at misinformation spread during a particular event.

I will leave you there and I will see you soon.