An analytical framework of Twitter analysis for wildfire hazards

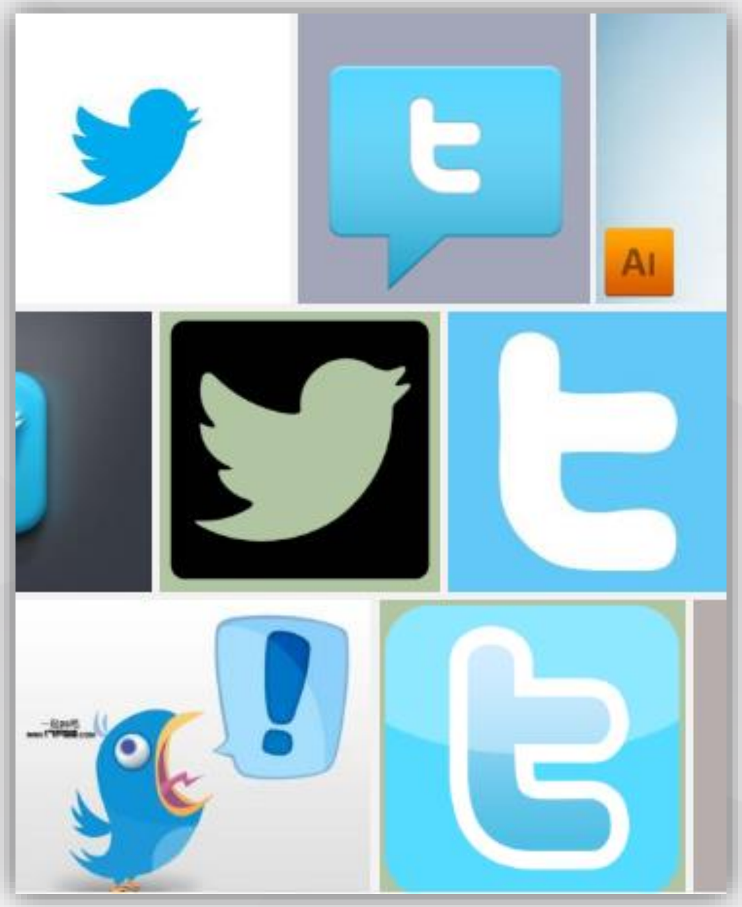
Human Dynamics and Big Data 2016

Xinyue Ye, Zheye Wang Department of Geography, Kent State University
Ming Tsou Department of Geography, San Diego State University

01 Introduction

As more and more fire-prone areas have been urbanized, people's livelihoods in the western USA have been severely influenced by the increasingly frequent wildfires.

1. Introduction



- Many efforts have been made to increase disaster-related information.
- Social sensing techniques featured by various big data sources such as **social media data** and **taxi trajectory data** are gaining increasing attention from domain scientists.
- Social media especially **Twitter** has been applied to "strengthen situational awareness and improve emergency response".

1. Introduction

1

2

3

4

wildfire exposure modeling

(Ager et al. 2014a, b; Thompson et al. 2015; Youssouf et al.2014)

wildfire risk assessment

(Chuvieco et al. 2010, 2012; Martı´nez et al. 2009; Padillaand Vega-Garcı´a 2011; Rodrigues et al. 2014)

wildfire and wildland–urban interface(WUI)

(Herrero-Corral et al. 2012; Massada et al. 2009; Schulte and Miller 2010)

wildfire–climate interactions

(Gillett et al. 2004; Liu et al. 2014; Westerling et al. 2006)

In order to achieve a better understanding of the occurrences and patterns of spread of wildfires, efforts by domain scientists have been made from various perspectives

1. Introduction

5

Wildfire management agencies have incorporated various wildfire detection systems, e.g., the general public, lookout towers, terrestrial mobile brigades, and aerial reconnaissance (Rego et al. 2013)

6

The Wildland Fire Decision Support System (WFDSS) has been developed (Calkin et al. 2011)

In order to achieve a better understanding of the occurrences and patterns of spread of wildfires, efforts by domain scientists have been made from various perspectives

1. Introduction

Space and time are strongly related to situational awareness in emergency events.

studies

- **De Albuquerque et al. (2015)**: carried out a spatial analysis and found a strong spatial relationship between locational proximity to floods and the usefulness of the messages for crisis management.
- **Guan and Chen (2014)**: found that the ratio of tweets associated with Hurricane Sandy to general tweets increased gradually before this disaster, peaked when it landed, and then gradually decreased
- **Huang and Xiao (2015)**: indicated that messages posted by Twitter users varied with the temporal process of a disaster and thus could provide useful information for improving situational awareness

1. Introduction

some studies focused on mining the actual content of social media messages to improve knowledge about disaster situations.

Qu et al. (2011)

divided the earthquake related microblog messages with valuable information for improving situational awareness into four categories.

Cameron et al. (2012)

developed a platform for emergency situation awareness, which could detect emergent incidents and classify tweets as interesting or not.

Imran et al. (2013a b)

Utilized machine learning methods to extract informative Twitter messages.

Imran et al. (2014)

further designed an Artificial Intelligence for Disaster (AIDR) platform.

1. Introduction

In disaster situations, people may also tend to obtain situational updates and gain situational awareness from the informative messages shared by opinion leaders.

studies

- **Cheong and Cheong (2011)**: found that local authorities, traditional media reporters, and, etc. are important players in spreading situational information during 2010–2011 Australian floods.
- **Kogan et al. (2015)**: indicated that local government authorities and the media are the most important nodes in the retweet network during the 2012 Hurricane Sandy.
- **Starbird and Palen (2010)**: A similar phenomenon was also observed

1. About this paper

This paper presents the findings from **examining the spatial and temporal variations of wildfire-related tweets** and from our attempt to **characterize wildfire by the discussion topics in the collected tweets**, as well as from investigating the **role of opinion leaders** in people's acquisition of wildfire-related information.

- **Introduce our data**
- **Related methodology**
- **Discuss the findings and their implications**
- **What future pursuits on this topic can be**

02 Data and methodology

2.1. Data

We used Twitter search API (https://search.twitter.com/) to collect wildfire-related Tweets. Our collection process included two phases.

- First, we collect any tweet that contained either of the two keywords—”fire” and ”wildfire”
- Second, we glean tweets associated with specific wildfires based on keywords which are places where wildfires occurred. The keywords were randomly selected from a list of places. (see Table 2)
- checking whether a “fire” or “wildfire” also appeared in the collected tweets.

2.1. Data

Table 2 Overview of the major wildfires in May, 2014. *Source*: complied from http://www.fire.ca.gov/

Major wildfires	Time of outbreak (UTC)	Time of 100 % contained (UTC)	Location	Long/lat	Acres
Bernardo Fire	May 13, 11:00	May 17, 20:14	Off Nighthawk Lane, southwest of Rancho Bernardo	−117.133/33.003	1548
Tomahawk Fire	May 14, 9:45	May 19, 9:20	Traveled from Naval Weapons Station, Fallbrook to Camp Pendleton	−117.285/33.353	5367
Poinsettia Fire (Carlsbad fire)	May 14, 10:30	May 17, 12:00	Off Poinsettia Ln & Alicante Rd in Carlsbad	−117.278/33.112	600
Highway Fire	May 14, 13:00	May 15, 18:30	Off Old Hwy 395 and I-15 in the Deer Springs area	−117.162/33.312	380
River Fire	May 14, 12:12	May 19, 9:20	North River Road and College Blvd., Oceanside	−117.747/33.251	105
Cocos Fire (San Marcos Fire)	May 14, 16:00	May 22, 18:15	Village Drive and Twin Oaks Road, San Marcos	−117.160/33.114	1995
Freeway Fire	May 14, 17:43	May 20, 11:30	Naval Weapons Station, Fallbrook	−117.260/33.370	56
Pulgas Fire	May 15 14:45	May 21, 17:00	Off Interstate 5 at Las Pulgas Rd, north of Oceanside	−117.463/33.303	14,416
San Mateo Fire	May 16, 11:24	May 20, 23:30	In the Talega area of Marine Corps Base Camp Pendleton	−117.300/33.286	1457

2.1. Data

Tweets collected in the first phase could be used in analysis of all dimensions (i.e., space, time, content, and network).

Tweets gleaned in the second phase are of particular importance for spatial analysis.

Our study period spans from May 13, 2014, when the first wildfire occurred, to May 22, 2014, when most of the destructive wildfires were 100 % contained. A radius of 40 miles was set to specify a circular area (centered at downtown) to cover the majority of San Diego County.

2.2. Methodology

Several specific methods were used in our study :

Kernel density estimation (KDE)

performed to analyze the spatial pattern of wildfire-related tweets

Text mining

identify conversational topics

Social network analysis

detect the opinion leaders in wildfire hazards

2.2. Methodology: KDE

- KDE imported the coordinates of tweets and exported a raster formatted map where each cell was assigned a value to represent the intensity level (Han et al. 2015).

- To deal with the impact of population, a dual kernel density estimation (Dual KDE) was employed.

Dual KDE Map =Each Cell Value of Tweets Map/Each Cell Value of Population Map

2.2. Methodology: Text Mining

A text mining for identifying important terms and term clusters in wildfire-related tweets.
Using the "tm" package in R 3.1.2.

FIRST

cleaned the raw tweets by removing URLs and stop words

SECOND

Obtained a term-document matrix, where a row stood for a term and a column for a tweet

THIRD

With k-means clustering method, terms which appeared frequently in the same document were grouped into one cluster.

03

Spatial and temporal analysis of wildfire Twitter activities

3. Spatial and temporal analysis of wildfire Twitter activities

First

Checked the temporal evolution of wildfire tweets and compared it with the wildfire's temporal evolution.

Second

Examined whether the impact areas are clusters of wildfire tweets or not.

we analyze the spatial and temporal relationship between social media activities and wildfire disruptions from the following two perspectives.

3. Spatial and temporal analysis of wildfire Twitter activities

- Table 2 demonstrates some basic spatiotemporal information of the major wildfires occurred in our study period.

Table 2 Overview of the major wildfires in May, 2014. *Source*: complied from http://www.fire.ca.gov/

Major wildfires	Time of outbreak (UTC)	Time of 100 % contained (UTC)	Location	Long/lat	Acres
Bernardo Fire	May 13, 11:00	May 17, 20:14	Off Nighthawk Lane, southwest of Rancho Bernardo	−117.133/33.003	1548
Tomahawk Fire	May 14, 9:45	May 19, 9:20	Traveled from Naval Weapons Station, Fallbrook to Camp Pendleton	−117.285/33.353	5367
Poinsettia Fire (Carlsbad fire)	May 14, 10:30	May 17, 12:00	Off Poinsettia Ln & Alicante Rd in Carlsbad	−117.278/33.112	600
Highway Fire	May 14, 13:00	May 15, 18:30	Off Old Hwy 395 and I-15 in the Deer Springs area	−117.162/33.312	380
River Fire	May 14, 12:12	May 19, 9:20	North River Road and College Blvd., Oceanside	−117.747/33.251	105
Cocos Fire (San Marcos Fire)	May 14, 16:00	May 22, 18:15	Village Drive and Twin Oaks Road, San Marcos	−117.160/33.114	1995
Freeway Fire	May 14, 17:43	May 20, 11:30	Naval Weapons Station, Fallbrook	−117.260/33.370	56
Pulgas Fire	May 15 14:45	May 21, 17:00	Off Interstate 5 at Las Pulgas Rd, north of Oceanside	−117.463/33.303	14,416
San Mateo Fire	May 16, 11:24	May 20, 23:30	In the Talega area of Marine Corps Base Camp Pendleton	−117.300/33.286	1457

3. Spatial and temporal analysis of wildfire Twitter activities

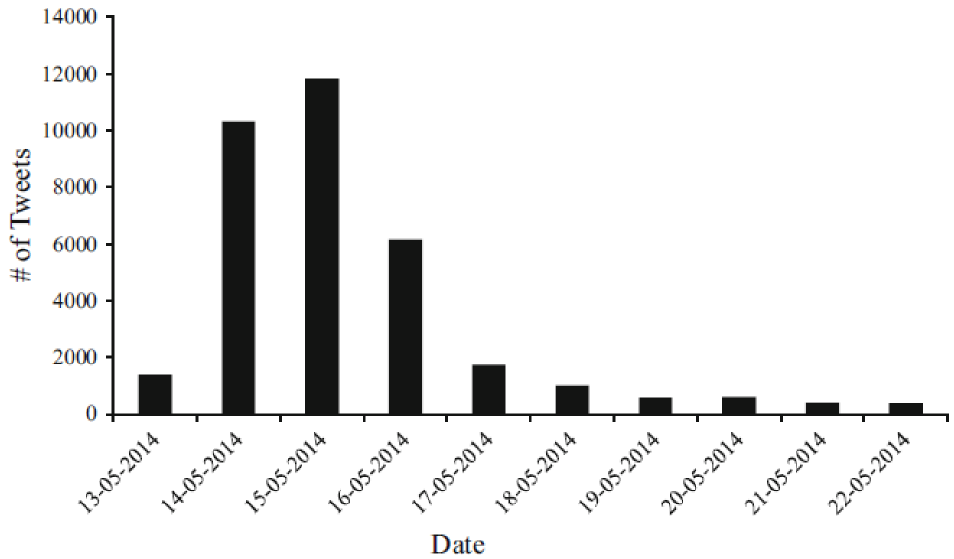


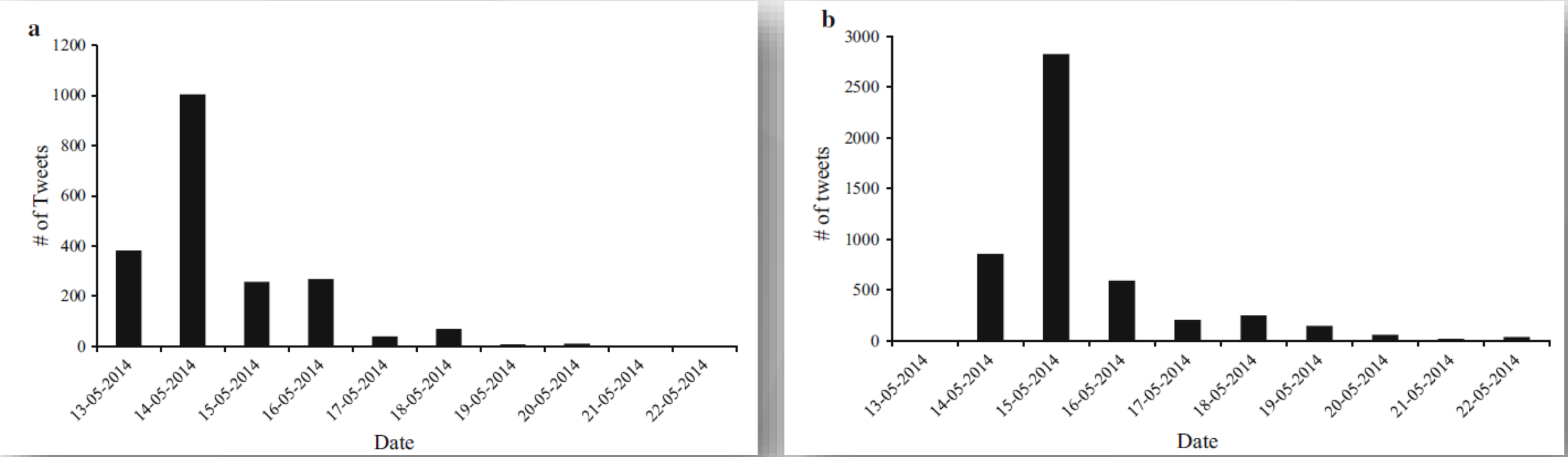
Fig. 1 Temporal evolution of wildfire-related tweets with keywords of "fire" and "wildfire"

Six of the nine wildfires occurred on May 14, which could explain why May 14 experienced a sudden increase in wildfire tweets (as shown by Fig. 1).

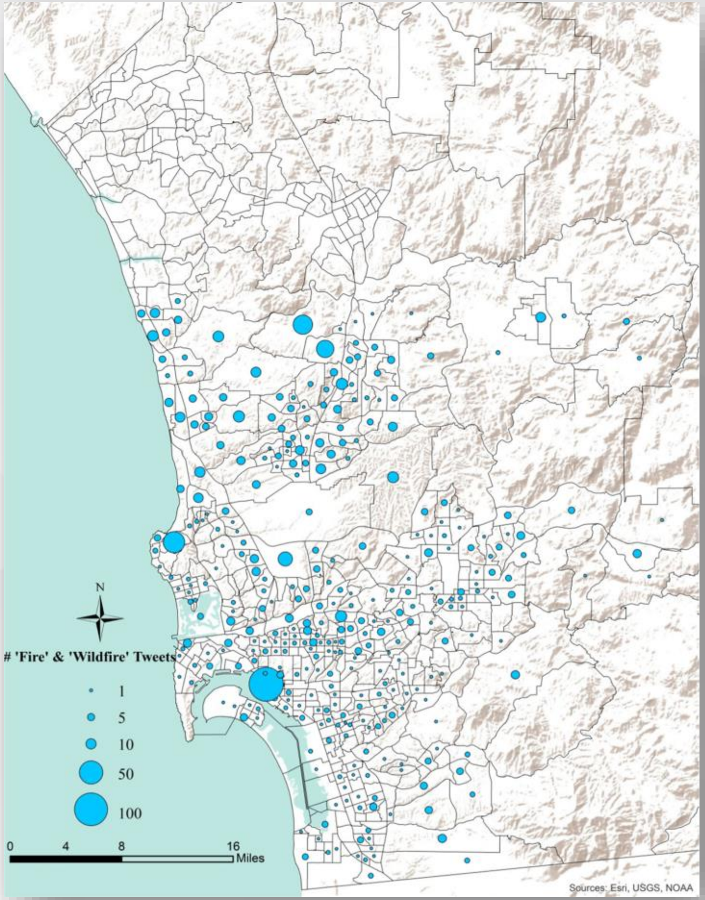
3. Spatial and temporal analysis of wildfire Twitter activities

- A temporally concurrent evolution of wildfire and its related tweets could also be observed from Fig. 2

- The Bernardo fire (a) and San Marcos fire (b) both had their corresponding tweets peak on the day after the breakout day. This 1-day time lag is probably because it takes time to spread information.

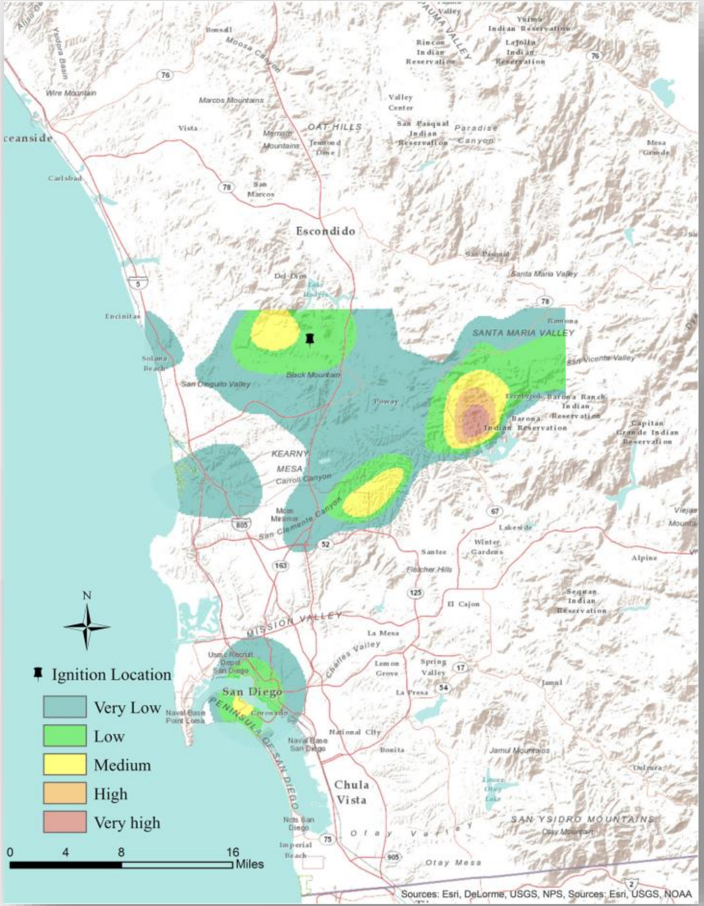


3. Spatial and temporal analysis of wildfire Twitter activities



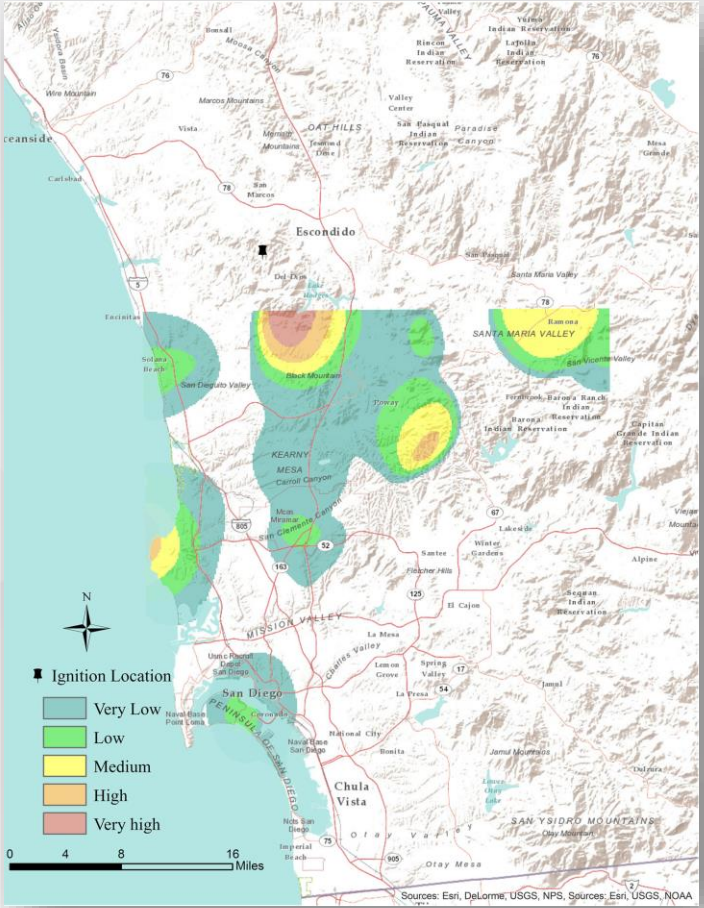
- Figure 3 shows that downtown area is the largest hot spot in terms of the number of "fire" and "wildfire" tweets.
- This may be due to the fact that a large population could generate numerous Twitter activities.

3. Spatial and temporal analysis of wildfire Twitter activities



To filter out the influence of population, dual KDE was performed to detect the clusters of tweets related to Bernardo fire and Cocos fire (see Figs. 4, 5 respectively)

3. Spatial and temporal analysis of wildfire Twitter activities

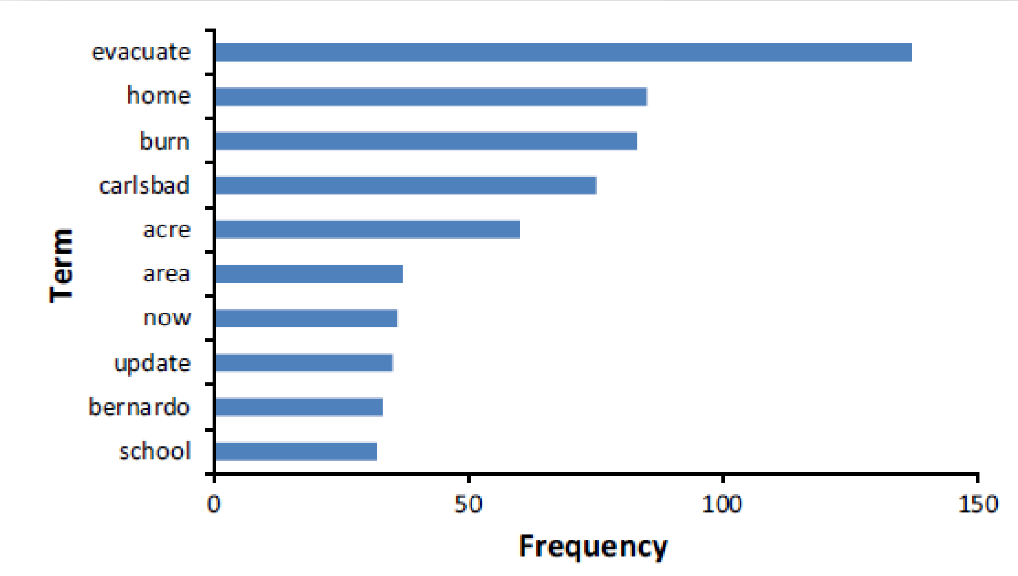


As shown by Figs. 4 and 5, the downtown area has become a low-value cluster, whereas clusters with values higher than medium are close to the wildfires' ignition locations

04 Topics and network

4. Topics and network

We first look at the importance of a term in tweets. Figure 7 shows us the top 10 frequent words. If a term appears frequently in tweets, it is regarded as important.



- the most important term is "evacuate", because the most urgent thing in wildfire situations is to evacuate

- a large part talked about the evacuation of homes, resulting in a high frequency of "home"

4. Topics and network

Table 3 shows the seven clusters, and within each cluster, only top three terms are shown. The number of clusters specified here is to ensure that we get the most but differentiated topics.

Number	Term clusters
Cluster 1	know; thank; firefight
Cluster 2	home; Carlsbad; burn
Cluster 3	wind; Carlsbad; area
Cluster 4	Carlsbad; contain; acre
Cluster 5	burn; evacuate; 4S Ranch
Cluster 6	acre; burn; contain
Cluster 7	evacuate; home; Bernardo

- cluster 1 stands for the topic related to thankfulness to firefighters
- cluster 2 is about the burned homes in Carlsbad
- cluster 3 is about the wildfire in Carlsbad area
- cluster 4 discloses a topic relevant to the containment percentage and impacted acres of Carlsbad wildfire

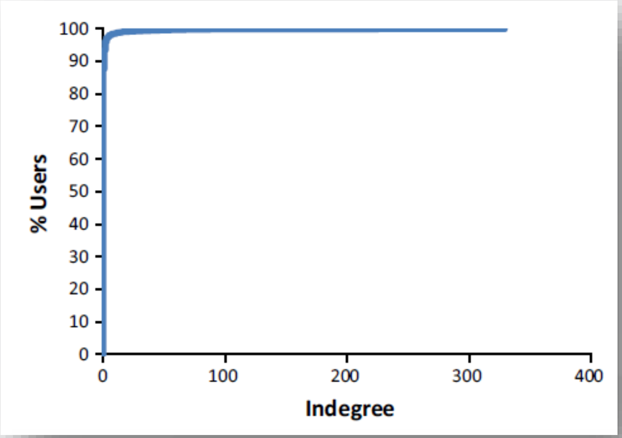
4. Topics and network

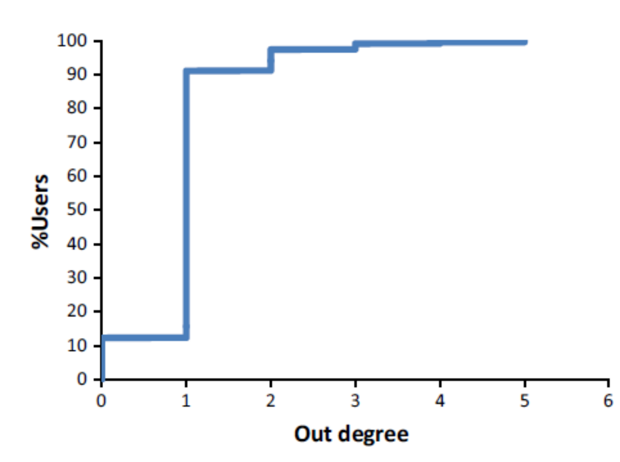
Table 3 shows the seven clusters, and within each cluster, only top three terms are shown. The number of clusters specified here is to ensure that we get the most but differentiated topics.

Number	Term clusters
Cluster 1	know; thank; firefight
Cluster 2	home; Carlsbad; burn
Cluster 3	wind; Carlsbad; area
Cluster 4	Carlsbad; contain; acre
Cluster 5	burn; evacuate; 4S Ranch
Cluster 6	acre; burn; contain
Cluster 7	evacuate; home; Bernardo

- cluster 5 represents the topic associated with the evacuation caused by a burning wildfire in 4S Ranch
- cluster 6 is a topic on damage report

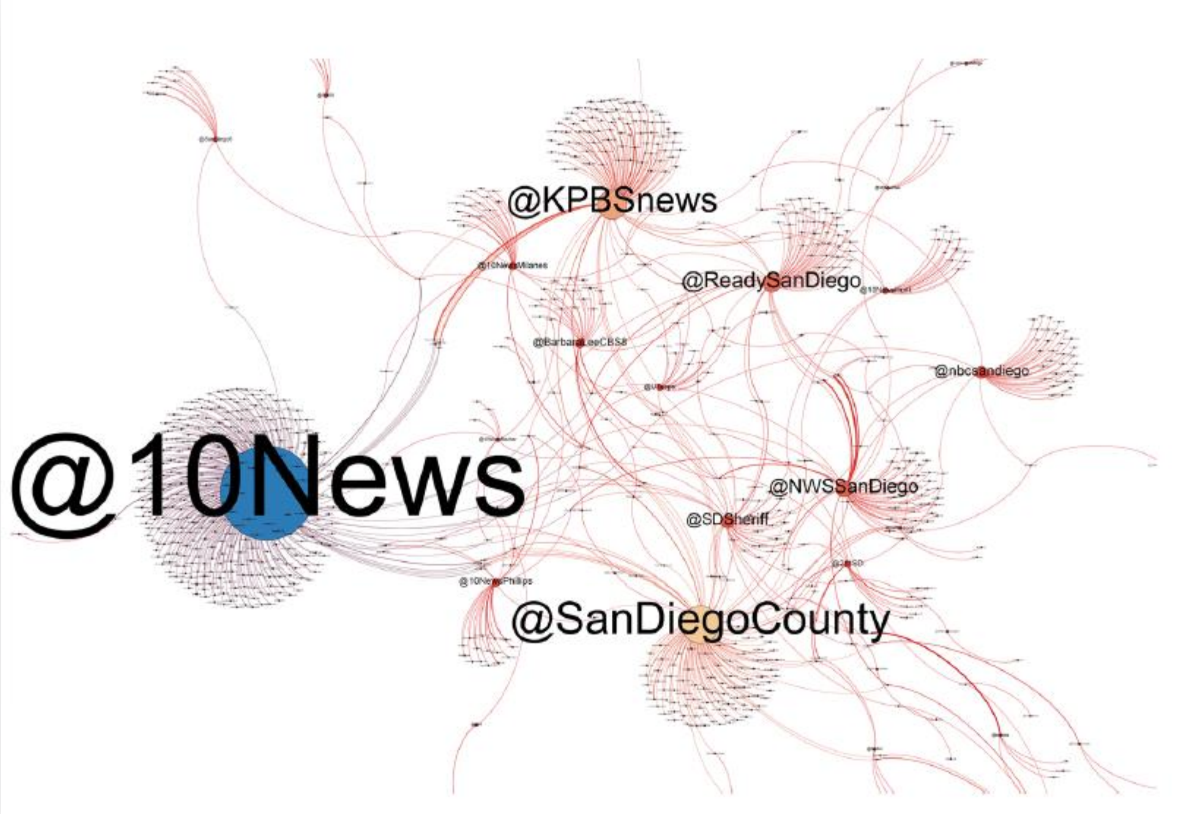
4. Topics and network





- The social network analysis was built based on the retweet relationship. We calculated the indegree and outdegree for each node.
- Figure 8 shows more than 85 % nodes had no users retweet their messages.
- Fig. 9 shows upward 90 % of users retweeted only one user or none.
- There are dominant users which act as hubs in the information exchange network.

4. Topics and network



The nodes of @10news, @KPBSnews, and @nbcsandiego are Twitter accounts owned by three local news media in San Diego.

05 Conclusion and discussion

5. Conclusion and discussion

Conclusion

spatial and temporal patterns of wildfire-related tweets

Our analysis confirmed a temporally concurrent evolution of wildfire and wildfire-related Twitter activities.

Mining topics can extract useful information

We found that people's geographical awareness is strong during emergency events

opinion leaders play an important role

We found that some elite users such as local authorities and traditional media reporters are dominant in the retweet network

simultaneous analysis

simultaneous analysis of the four dimensions might be able to provide some new insights

drawbacks

First

although the searching range could cover the majority of San Diego County, some places where wildfire occurred were not contained.。

Second

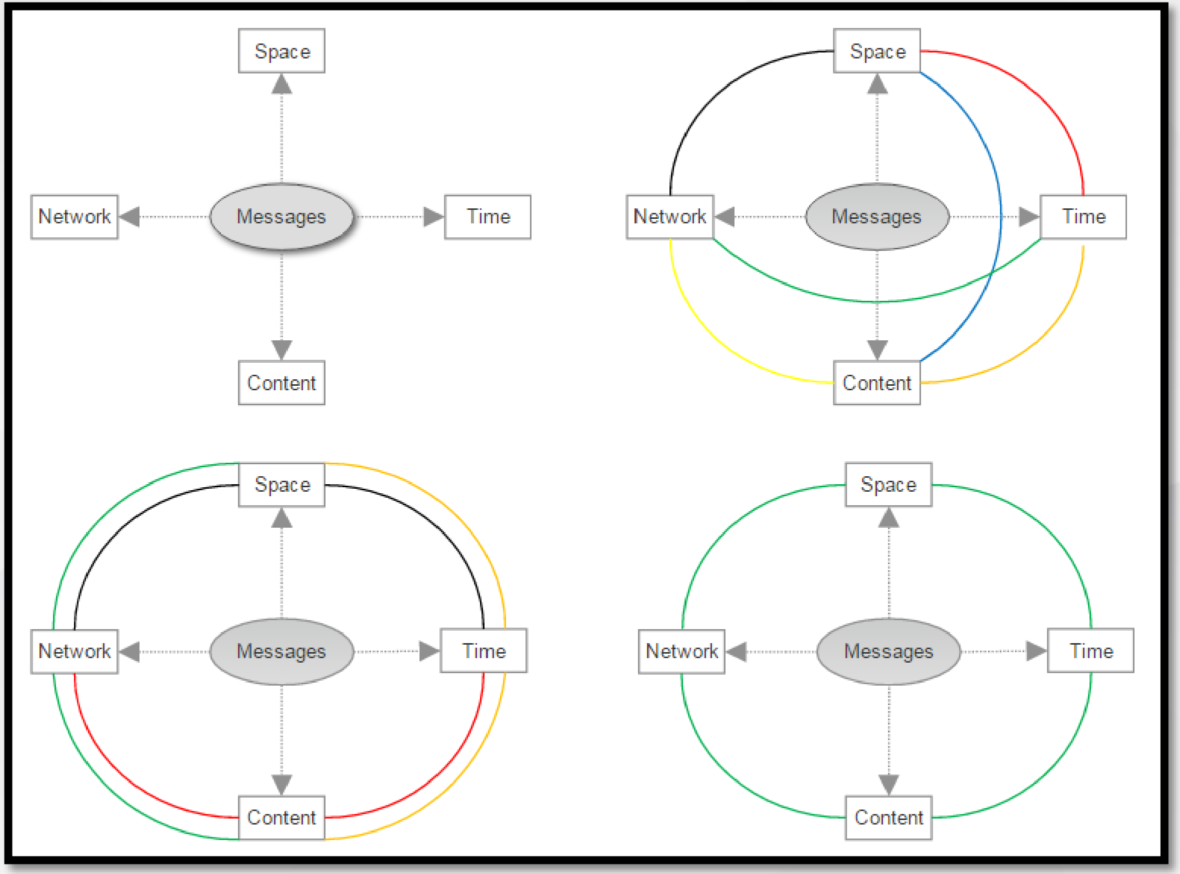
the 1% sample limitation may lead to question that whether the sampled data are a valid representation of the overall wildfire Twitter activities

Third

the social network in our research is only based on the retweet relationship, while other types of could be used in future study.

Fourth

the social network analysis centered on the investigation of opinion leaders in wildfire situation and thus overlooked the information diffusion process including its components, phases, and characteristics.



Four dimensions have 15 possible combinations

$$(C_4^1 + C_4^2 + C_4^3 + C_4^4)$$

THANKS