11-2013

Social sensing for urban crisis management: The case of Singapore Haze

Philips Kokoh PRASETYO
Singapore Management University, pprasetyo@smu.edu.sg

Ming GAO
Singapore Management University, mgao@fudan.edu.cn

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

Christie N. SCOLLON
Singapore Management University, cscollon@smu.edu.sg

Citation

PRASETYO, Philips Kokoh; GAO, Ming; LIM, Ee Peng; and SCOLLON, Christie N.. Social sensing for urban crisis management: The case of Singapore Haze. (2013). *Social Informatics: 5th International Conference, SocInfo 2013, Kyoto, Japan, November 25-27, 2013: Proceedings*. 8238, 478-491.
Available at: https://ink.library.smu.edu.sg/sis_research/1964

Social Sensing for Urban Crisis Management: The Case of Singapore Haze

Philips Kokoh Prasetyo, Ming Gao, Ee-Peng Lim, and Christie Napa Scollon

Singapore Management University
{pprasetyo,minggao,eplim,cscollon}@smu.edu.sg
http://www.larc.smu.edu.sg

Abstract. Sensing social media for trends and events has become possible as increasing number of users rely on social media to share information. In the event of a major disaster or social event, one can therefore study the event quickly by gathering and analyzing social media data. One can also design appropriate responses such as allocating resources to the affected areas, sharing event related information, and managing public anxiety. Past research on social event studies using social media often focused on one type of data analysis (e.g., hashtag clusters, diffusion of events, influential users, etc.) on a single social media data source. This paper adopts a comprehensive social event analysis framework covering content, emotion, activity, and network. We propose a set of measures for each dimension accordingly. The usefulness of these analyses are demonstrated through a haze event that severely affected Singapore and its neighbors in June 2013. The analysis, conducted on both Twitter and Foursquare data, shows that much user attention was given to the haze event. The event also saw substantial emotional and behavioral impact on the social media users. These additional insights will help both public and private sectors to prepare themselves for future haze related events.

1 Introduction

1.1 Motivation

Social sensing is an important step towards understanding how a disaster or social event affects individuals and communities. A good understanding of the event allows us to answer questions about its severity as well as the effectiveness of interventions introduced during the event. For example, during an earthquake, people may suffer from unsafe environment, inadequate supply of food and water, loss of houses, missing family members and lack of medical care. This wide range of concerns naturally become the topics of discussion among people, as well as the focuses of disaster management and rescue efforts.

In the past, social events are mostly reported by news media. One therefore can only perform the event impact assessment by analyzing news content or conducting field surveys on the event-affected people. Both approaches unfortunately require much time and resources. They also could not capture a complete view of the event due to limited reach to the larger user population.

A. Jatowt et al. (Eds.): SocInfo 2013, LNCS 8238, pp. 478–491, 2013.

With the popularity of social media, there is a large amount of user-generated content that can be harnessed for event analysis. Unlike traditional news media, social media content is generated directly from the user population and therefore offers direct access to both individual and community levels feedbacks. Other than its textual content, social media also records user behaviors that can be very useful in event analysis. For example, users may express their opinions online using votes and ratings, share interesting online content with their friends, etc. These non-textual behavioral data can be used to determine user activity patterns during the event.

1.2 Objective

In June 2013, Singapore experienced the worst haze in its history. The haze was a unique event that affected all people in Singapore. At the time, local media covered almost nothing but haze for several days. Office chatter and neighborly greetings were abuzz with talk of the haze. Our observations suggested that haze was on everyone's mind. But is there a way to measure the consciousness of Singaporeans during the haze to know their thoughts and feelings? Is there a way to quantify the impact of the haze on human activity?

To answer the above questions, we study social media usage by Singapore users during the haze event so as to derive some insights about peoples reactions to the haze. We adopt a social analysis framework that consists of four types of analysis on the event-relevant social media data (see Figure 1). The analysis can also conducted on data divided into different time intervals for trend analysis. The analysis techniques adopted are:

- *Content analysis*: This includes analysis on all textual social media data. The purpose is to determine the content topics, content objects (e.g., photo images, videos), representative keywords or keyphrases that help to explain the event. Content topics can be derived by clustering words or assigning them with topic labels [1].
- *Emotion analysis*: One can analyze tweet content for different emotion keywords to determine the state of user emotions as they generate the tweets. Bollen, Mao and Pepe analyzed user mood on Twitter in six dimensions using a set of words for each mood dimension [2]. In this paper, we use a similar approach using selected emotion words from Pennebaker's *Linguistic Inquiry and Word Count* (LIWC) full dictionary [3].
- *Activity analysis*: The activity behavior of users can be determined by activity words mentioned in the social media content, or by observing their actions recorded in social media data. In this paper, we use Foursquare check-in's to determine the places visited by social media users. When the activity data are geo-coded, one can even determine the locations of activities [4].
- *Network analysis*: Network analysis focuses on constructing human networks, and analyzing for each network central nodes (e.g., influential users, information gatekeepers), relationships (e.g., strong and weak ties) and communities [5]. Using network analysis, one can study how an event affects the network properties and dynamics (e.g., diffusion of information).

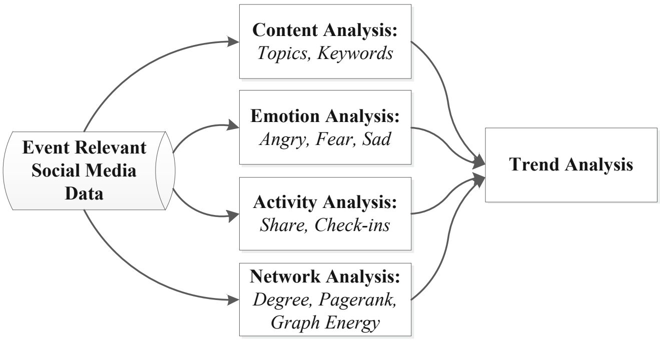


Fig. 1. Social Impact Analysis Framework

In our study, we observed that:

– The haze event attracted much attention from Singapore users only after the latter realised that the event lasted longer than expected. Substantially more people decided to tweet during the haze event.
– Users monitored the haze condition closely and depended a lot on traditional news media and government agencies for information. Nevertheless, they also demonstrated more negative emotion during the haze.
– Users reduced their outdoor activities causing fewer visits to eating places. This suggests that food and other businesses were quite badly affected by the haze event.
– The National Environment Agency (NEA) of Singapore emerged to be a central node in the network analysis during the event. Most traditional news media accounts also benefited from haze by seeing their centrality ranks improved. We however could not find any individual user gaining significant centrality rank.

1.3 Dataset Construction

The social media data used in this study are collected from Twitter and Foursquare. We collected Twitter data generated by about 130K public user accounts with Singapore stated as their user locations in June 2013. From this dataset known as SGTwitterData, we selected those that contain one of the following keywords: *sghaze*, *haze*, *mustbehaze*, and *blamethehaze*. This selected subset of the data is called HazeTwitterData. Using carefully selected keywords to collect event relevant tweets was also used in other works [6,7].

We also collected a month of FourSquare check-ins data in June 2013 which were generated by the same set of Singapore users. We call this data the 4SQ-Data. Due to some crawling problem, we were not able to gather complete set of tweets on 13 June 2013. Hence, we would leave that day out of our study below. We did not include private user accounts as their profile and tweet information are not open to public.

1.4 Paper Outline

The rest of this paper is organized as follows. Section 2 describes some previous work regarding social sensing, and social sensing for crisis management. Our analysis on the content and activity dimensions of the data is presented in Section 3. More detailed analysis on emotional states is discussed in Section 4. Section 5 shows the dynamics of Singapore Twitter network during the haze crisis. Finally, we conclude our study in Section 6.

2 Related Work

Using hashtags that have been widely used to annotate tweets, Lehmann et al. studied the clusters of hashtags and their evolution over time and found four types of events: (i) those that attract attention before and during peak (measured by number of mentioned tweets); (ii) those that attract attention during and after peak; (iii) those that attract attention symmetrically around peak; and (iv) those that attract attention on a single day of the peak [8]. Crises are likely of type (ii) due to its unexpectedness and social impact.

Earle, Bowden and Guy found that sensing Twitter for earthquake events allows one to detect earthquake with human impact early among many earthquakes that have actually happened, especially in regions where the siesmic sensors are not available [7]. The tweet content also provides very good contextual insights into the earthquake events.

In [6], the dissemination of rumors and news on Twitter during the 2010 earthquake in Chile was analyzed. Rumor spreading and news sharing are user behaviors prevalent in crisis events. The work analyzes about 4.7M tweets from 716K users during the event. It was found that the earthquake related content propagated very quickly on Twitter. Rumors were also found to propagate (or be retweeted) very differently from news as they are more likely to be refuted and questioned by users.

Cheong and Cheong conducted social network analysis on Twitter data related to Australia flooding events in 2010 and 2011 [9]. Two social networks were used, namely a retweet/mention network and a user-URL network. The high degree nodes in the two networks represent the *influential users* and *popular resources* respectively. These two sets of nodes lead us to find the users active and useful content in the events. Nevertheless, it also pointed out the local authorities did not manage to play influential nodes in the events.

To help emergency event-affected users to share tweets about the events, a Tweak the Tweet (TtT) syntax was proposed by Starbird and Stamberger to introduce a set of hashtags (e.g., #need, #offer, #iamok, etc.) to be used in tweets reporting the events [10]. Starbird and Palen found that very few users on the ground adopting the TtT syntax in the 2010 Haiti earthquake event but also several other users volunteering efforts to translate original tweets into ones that follow the TtT syntax [11]. These are examples of volunteerism and self-organizing behaviors that can be observed during crisis events. In the Singapore haze event studied in this paper, we unfortunately could not find tweets following the TtT syntax.

Compared with the above works, this paper adopts a comprehensive social analysis framework that covers the *content*, *emotion*, *activity* and *network* aspects. While the measures defined for each aspect are not new, there has not been any work to apply them all to a large-scale event such as the Singapore's haze event. The haze event involves haze-relevant hashtags and keywords that are of type-(ii) as most Singapore users did not expect it to happen. Unlike other disastrous events, telecommunication and transportation networks were not affected in the haze event. Users therefore can share their social media data and communicate in their usual ways.

3 Content and Activity Analyses

3.1 Overall Tweet Trend

Our first analysis tries to find out how the haze has affected the usage of social media. More than 26 million tweets are generated in June 2013 by about 130K public users in SGTwitterData. We measure the number of tweets in SGTwitterData each day. Figure 2 shows the daily tweet count of SGTwitterData in June 2013. As shown in the figure, the daily tweets generated by Singapore users surged on 19 and 21 June when Pollutant Standards Index (PSI) hit peak numbers at 321 and 400 respectively[1]. The surge on 17 June was relatively small. This suggests that there were much more Twitter data generated by Singapore users during the haze crisis. Thus, the haze crisis indeed has affected the usage of social media.

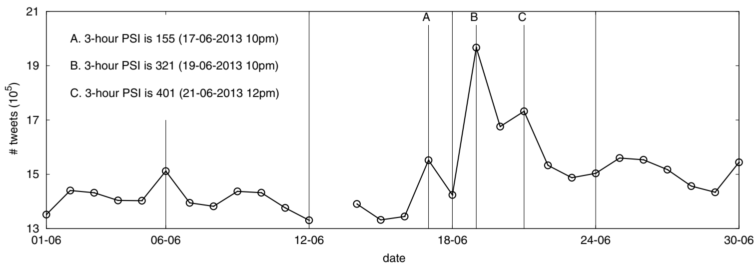
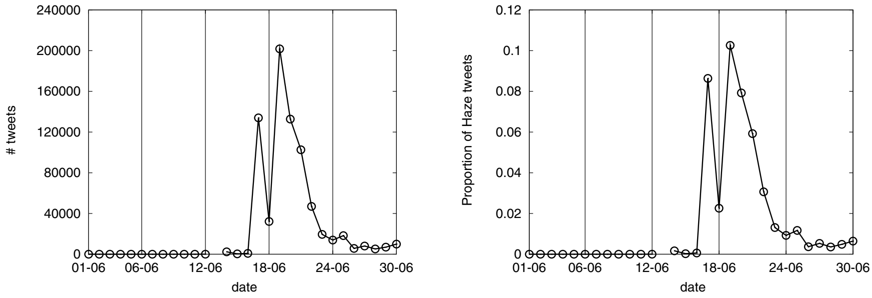


Fig. 2. Daily Tweet Count

3.2 Haze-Related Tweet Trend

Analysis on HazeTwitterData was performed to learn how big was the impact of haze crisis in the overall twitter activity. Daily tweet count of HazeTwitterData is shown in Figure 3(a). The figure shows that the daily tweet count of

[1] According to NEA, PSI reading beyond 100 is considered unhealthy.



(a) Daily Haze-related Tweet Count (b) Proportion of Haze-related Tweet

Fig. 3. Daily Haze-related Tweet

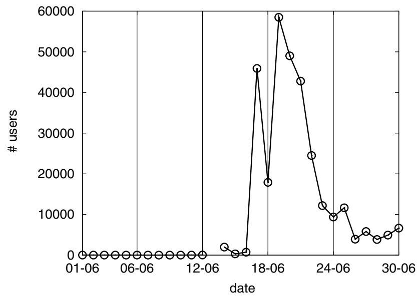


Fig. 4. Daily User Count

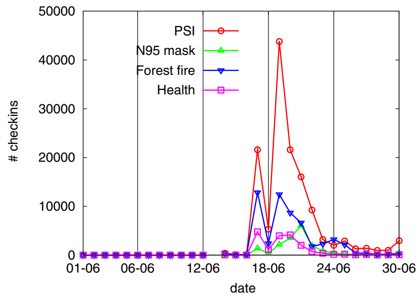


Fig. 5. Topic Trend

HazeTwitterData surged on 17 June. There are almost no tweets about haze before 17 June. The number of haze related tweets reduced on June 18 possibly due to the common belief that the haze would not last long. On 19 June onwards, many more haze related tweets were generated as users realized that the haze problem was worsening and would last for a longer period. After 21 June when the haze began to subside, the number of haze-related tweets decreased substantially, but remained higher than the number before the haze.

Daily proportion of HazeTwitterData is displayed in Figure 3(b). The figure shows that on 17 and 19 June, HazeTwitterData accounts for about 8% and 10% of all tweets generated by Singapore users. After 19 June, the proportion of haze related tweets continued to stay substantially higher than the proportion before the haze crisis.

Daily user count in HazeTwitterData can be seen in Figure 4. Very few users mentioned haze-related contents before the number surged on 17 June. The number hit the highest number of users on 19 June with more than 58,000 users. Although the number decreases substantially after the haze subsided, few thousand of users still mentioned haze-related contents until the end of the month.

3.3 Topic Analysis

Analysis on tweet content reveals few popular topics in haze-related contents. From the top frequent words that appear in HazeTwitterData, we manually categorized the words into four topical categories, namely:

- PSI category: “nea”, “psi”
- N95 mask category: “mask”, “n95”
- Forest fire category: “forest”, “fire”, “Indonesia”, “Malaysia”, “Sumatra”, “nature”, “smoke”, “burn”
- Health category: “asthma”, “breathing”, “health”, “hospital”, “clinic”, “doctor”, “sick”, “respiratory”

Figure 5 shows that the overall trends of haze-related tweets under the above four topics follow that of overall haze-related tweets. The PSI reading captured the most user attention, and most tweets were about the PSI. The second largest topic was about forest fire and followed by the health topic. The N95 mask topic became more popular than the health topic on 21 June because of the high demand of N95 mask after the air quality reached its worst on 21 June.

3.4 Information Sources

URLs mentioned in tweets indicate external sources which bring in the information to social media. We examined the highly mentioned and retweeted domain names in HazeTwitterData to identify the information sources people trusted during the haze crisis. There were 61,889 tweets, and 36,312 retweets that contain URL in HazeTwitterData. The domain names were categorized into three categories, namely News, Government and Others.

Tables 1 shows that the top domain names mentioned and retweeted by users in June 2013 are mainstream news sites, a government agency, and several other popular social media sites. The domains marked with * are ones that appear in only one of two tables. The tables indicates that most users still referred to official news channels for information about haze.

Among the top domain names, NEA was the only government agency appeared to be an important source during the haze crisis. The NEA played an important role disseminating haze information to the public. The NEA published the 3-hour PSI reading every one hour. To observe how quickly the information got disseminated, we counted the number of retweets mentioning nea.gov.sg, and divided them into six disjoint time windows of ten minutes each across time. Figures 6 shows the retweet count of each time window on 17, 19, 20, and 21 June 2013 respectively.

The peaks in the figures indicate that most people responded (by retweeting the URL) to NEA announcements within ten minutes. This shows that people tracked NEA announcements closely. Furthermore, this also indicates that many people checked the PSI reading obsessively since most of NEA announcements were on PSI reading.

Table 1. Top Domain Names in HazeTwitterData in June 2013

(a) Top 10 Mentioned Domain Names

Rank	Domain	Category	#tweets
1	straitstimes.com	News	13,776
2	instagram.com	Others	10,460
3	nea.gov.sg	Government	7,240
4	todayonline.com	News	6,661
5	channelnewsasia.com	News	3,589
6	facebook.com	Others	2,607
7	youtube.com	Others	2,528
8	twitpic.com	Others	2,044
9	ask.fm*	Others	1,230
10	stomp.com.sg	Others	1,150

(b) Top 10 Retweeted Domain Names

Rank	Domain	Category	#retweets
1	straitstimes.com	News	11,193
2	nea.gov.sg	Government	6,421
3	todayonline.com	News	5,802
4	channelnewsasia.com	News	2,369
5	twitpic.com	Others	1,683
6	instagram.com	Others	1,543
7	youtube.com	Others	1,533
8	stomp.com.sg	Others	961
9	facebook.com	Others	922
10	yahoo.com*	News	367

3.5 Check-Ins

To study the impact of haze to user activities and businesses, we analyzed 4SQ-Data. Figure 7 shows Foursquare check-ins trend in June 2013 by Foursquare venue category. We observed four categories of check-ins:

- Food: restaurant, food, and cafe
- School: university, school, and college
- Shop: mall, shop, and department store
- Healthcare: hospital, clinic, doctor, pharmacy, drug store

The haze reduced Foursquare check-ins especially on 19-22 June 2013 (up to one day after the worst haze day). There was clear evidence that visits to shops and eating places were reduced substantially, by 20% to 50% respectively, during the three days (17, 19 and 21 June) that witnessed record breaking PSI values. Daily check-ins pattern of both shops and eating places were similar because many shops and eating places shared the same building. The reduction of check-ins to schools appeared to be less obvious. This could be due to school holidays during the haze period. At the same time, the number of check-ins to healthcare places was also reduced. However, healthcare locations typically make up a very small proportion of FourSquare check-ins in general.

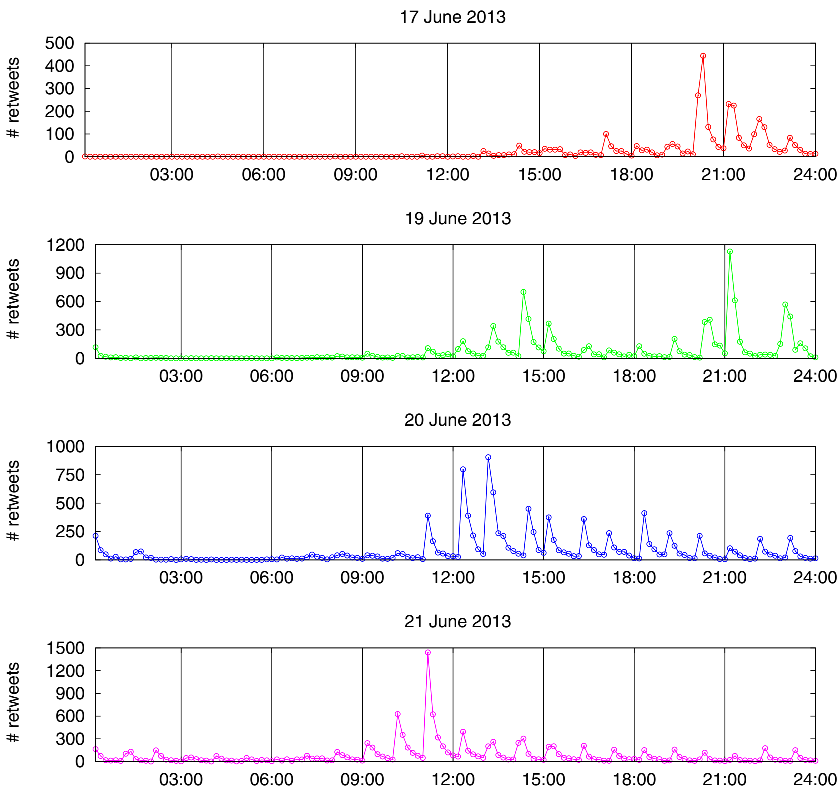


Fig. 6. Retweets of NEA Announcements

4 Emotion Analysis

Another important aspect during the haze crisis was emotion states of the people in Singapore. In this section, we present our analysis in sensing emotion state of users in Singapore during the haze crisis.

4.1 Types of Emotion

We examined the emotion state of the users by classifying the tweets into the categories below according to the emotion associated keywords from the LIWC full dictionary [3].

- Negative categories:
  - Anxiety category: "worried", "worry", "worries", "fear", "afraid", "frightened", "scared", "stress", "upset", "nervous", "anxious", "alarm", "tense", "distress", "panic", "die"
  - Anger category: "mad", "frustrate", "irritate", "annoy", "hate", "kill", "piss", "mean", "hostile", "disgust"

- Swear category: "piss", "fuck", "damn", "shit", "crap", "oh no", "OMG", "holy _", "FML"
- Low arousal negative category: "unhappy", "miserable", "sad", "depress", "hopeless", "gloomy", "tired", "sleepy", "lethargic", "fatigue", "helpless", "down", "dejected"

- Gratitude category: "thank", "thankful", "grateful", "blessed", "lucky", "fortunate", "pleased"
- Positive category: "happy", "love", "glad", "pleased", "relax", "calm", "relieve", "phew", "inspire", "proud", "joyful", "excite", "admire", "cheerful", "delight", "eager", "elate", "enthusiastic", "interest", "peaceful", "pleasant", "respect"

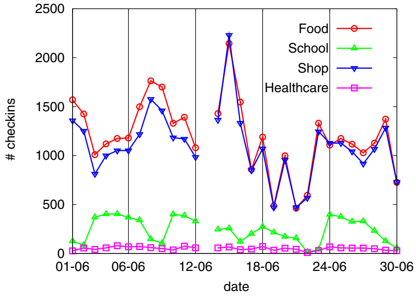


Fig. 7. Daily Foursquare Checkins

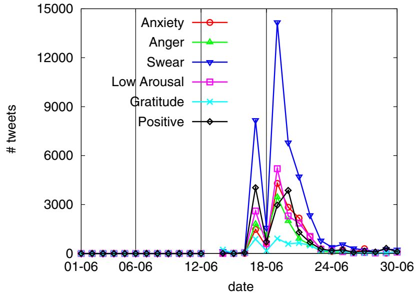


Fig. 8. Emotion Trend

4.2 Emotion Results

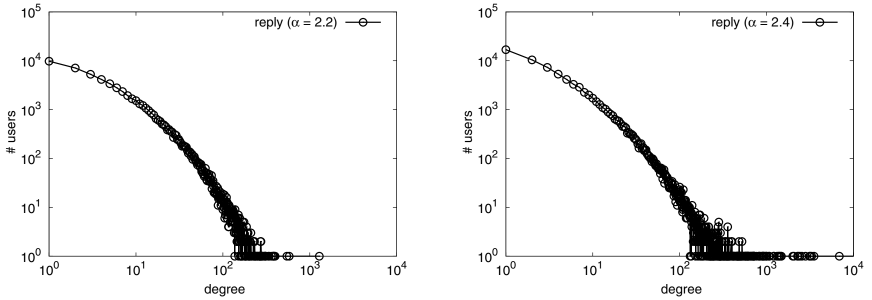
We observed that the negative emotions, especially the swear words, reached high peaks during the haze event as shown in Figure 8. Almost all negative emotion words, i.e., anxiety, anger, swear and low arousal words, hit the highest numbers of tweets on 19 June. There were more positive emotions expressed on 20 June before they dropped when the PSI index hit the record breaking 401 on 21 June. Grateful emotions didn't change much throughout the month.

The swear words were found about 6% and 7% of haze related tweets during the two peaks of haze event on 17 and 19 June respectively. After that, the percentage of swear words reduced to values less than 6%. On the worst haze days, i.e. 19 and 21 June, swear, low arousal, and anxiety words were the top three emotion expressed by Singapore Twitter users.

5 Network Analysis

5.1 Retweet and Reply Networks Construction

Based on HazeTwitterData, we construct a *reply network* and another *retweet network* for each day. An edge (u, v) is formed in the daily reply network if user u replies at least a tweet from user v, or user v replies at least a tweet from user



(a) Reply Network on 19 June 2013 (b) Retweet Network On 19 June 2013

Fig. 9. Degree distributions of Reply and Retweet Networks

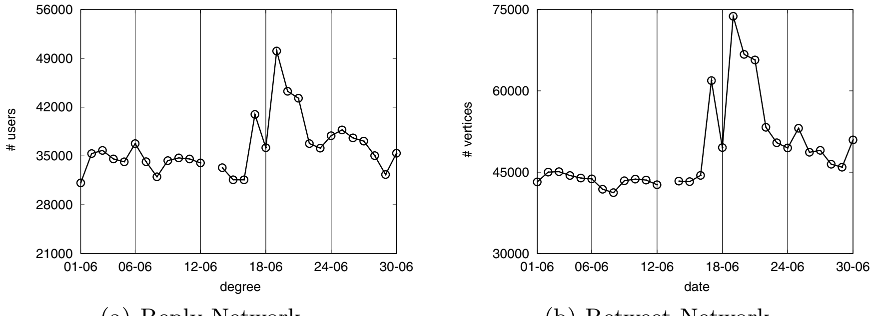
u in the day. The edges in the daily retweet network are created in a similar manner.

By checking the degree distributions of reply and retweet networks, we can conclude that they are quite similar to scale-free networks, where the scaling parameters α are estimated by the approach presented in [12]. This is illustrated by the degree distributions of reply and retweet networks on 19 June 2013 as shown in Figure 9.

5.2 Network Robustness Analysis

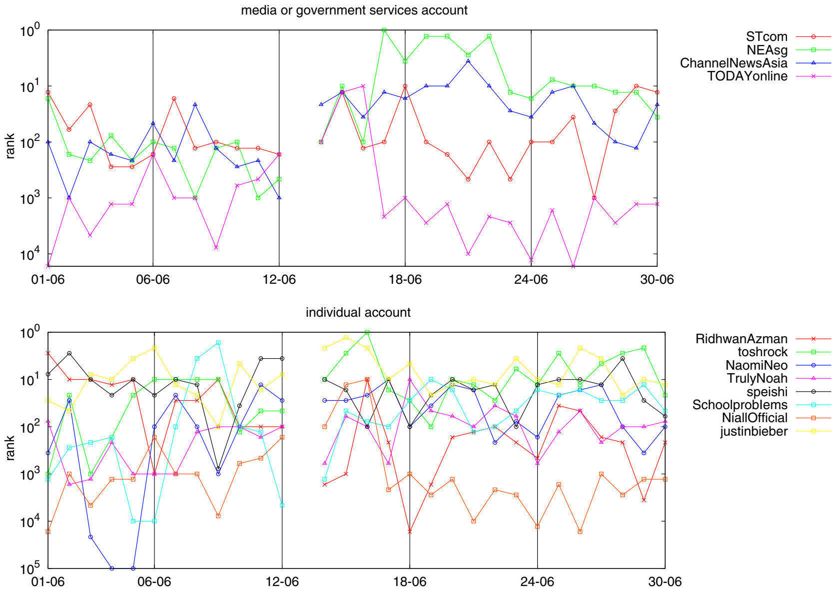
Network robustness determines how well its vertices are connected to one another so as to keep the network strong and sustainable. Larger network may be more robust as it is hard to change. Since the largest CC (Connected Component) is a good representation of the whole network, the size of the largest CC is also a simple measure to evaluate the robustness of a network.

Figure 10 shows the sizes of the largest CCs on both reply and retweet networks from 1st June to 30th June (with the exception of Jun 13th when we experienced a data crawling problem). We observe that: (1) many users involved

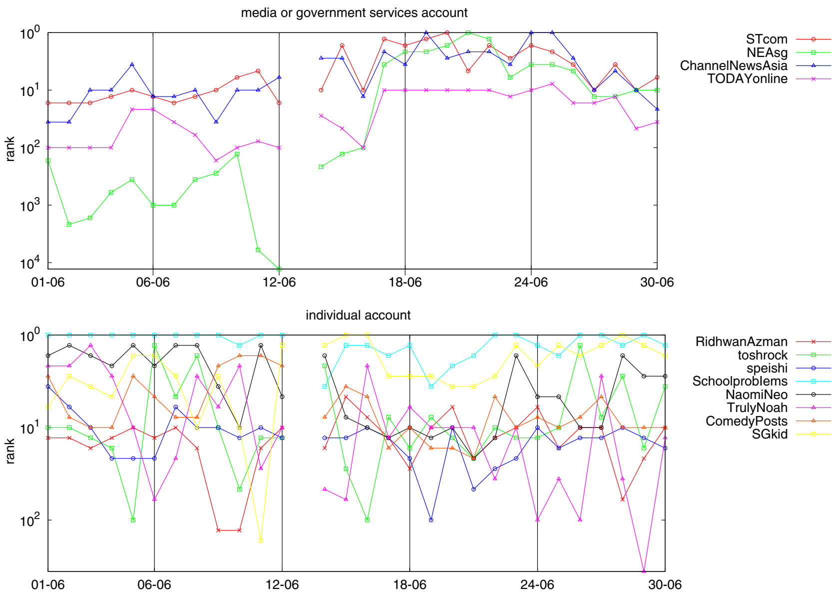


(a) Reply Network (b) Retweet Network

Fig. 10. The Sizes of the Largest CCs on Reply and Retweet Networks



(a) Reply Network



(b) Retweet Network

Fig. 11. Influential Users in terms of PageRank Centrality on Reply and Retweet Networks

in replying one another during the crisis, even after the crisis; (2) users were also more likely to retweet with one another during the crisis.

5.3 Centrality Analysis

Another natural question is who were the influential users and popular sources within the reply and retweet networks during the event. We therefore employ centrality measures to answer this question as the centrality of a vertex determines its relative importance within a network. Multiple centrality measures, such as degree centrality, pagerank centrality, betweenness centrality, closeness centrality, and eigenvector centrality, are widely used in network analysis [13,14]. Note that more inferential users are more likely to share, diffuse and propagate information on Twitter platform. Meanwhile, taking in account the scalability of measures, we employ the pagerank centrality, to determine the influential users and popular sources from the interaction networks.

Figure 11 shows the influential users and popular sources, measured by pagerank centrality, on the reply and retweet networks, where each curve represents the daily pagerank (in log scale) of a user from 1st June to 30th June. We find that news mediums or government services, such as *STcom*[2], *NEAsg*[3], *ChannelNewsAsia*[4], and *TODAYonline*[5], became popular. However, influential individual users are always on the top list regardless of whether there is an event. We can conclude that government services played the most important role during the crisis, followed by the new media. Except June 18, an interesting observation is that the NEA Twitter account (*NEAsg*) attracts more attention during the event, even after the event. This is due to the common belief on June 18 that the haze would not last long.

6 Conclusion

We may think of social media as a modern frivolity mainly to be used as a source of fun, but it can also supply some useful and quantifiable information. Twitter provides a window into the stream of consciousness of Singaporeans like no other technology at present can. Even large-scale surveys cannot track responses to an event as close in time to the occurrence or with such large samples. Geo-location data from FourSquare can quantify real activity to confirm or disconfirm personal observations. Our analysis of Twitter feeds found the impact of the haze on peoples lives was undeniable and intense. They drastically reduced their activity to food and shopping venues until the haze cleared. Instead of outdoor activities, people responded by turning to social media to express themselves. Their expressions were primarily ones of shock, anger, and other negative emotions. At the same time, people relied heavily on official sources of information about

[2] The Straits Times.
[3] NEA account.
[4] Official site of Channel NewsAsia.
[5] Singapore's most popular compact newspaper.

the haze, and they used social media to spread this information. More details about this study can be found at http://research.larc.smu.edu.sg/sghaze.

Acknowledgement. This work is supported by the National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, and National Research Foundation (NRF).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
2. Bollen, J., Mao, H., Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: ICWSM. (2011)
3. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates (2001)
4. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An empirical study of geographic user activity patterns in Foursquare. In: ICWSM. (2011)
5. Newman, M.E.J.: The structure and function of complex networks. SIAM Review **45** (2003) 167–256
6. Mendoza, M., Poblete, B., Castillo, C.: Twitter Under Crisis: Can we trust what we RT? In: Proceedings of the first workshop on social media analytics
7. Earle, P.S., Bowden, D.C., Guy, M.: Twitter earthquake detection: earthquake monitoring in a social world. Annals of Geophysics **54**(6) (2012)
8. Lehmann, J., Gonçalves, B., Ramasco, J.J., Cattuto, C.: Dynamical classes of collective attention in twitter. In: WWW. (2012)
9. Cheong, F., Cheong, C.: Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods. In: PACIS. (2011)
10. Starbird, K., Stamberger, J.: Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting. In: ISCRAM. (2010)
11. Starbird, K., Palen, L.: Voluntweeters: Self-organizing by digital volunteers in times of crisis. In: CHI. (2011)
12. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Review **51** (2009) 661–703
13. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry (1977) 35–41
14. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the Web. (1999)