Visualizing Research: The Spread of Rumors on Twitter

Cynthia Andrews
COM
Seattle, United States
caandrew@uw.edu

Zeno Koller
CSE
Seattle, United States
kollerz@uw.edu

Rama Gokhale
CSE, INFO
Seattle, United States
ramag@uw.edu

Graeme Britz
CSE
Seattle, United States
grbritz@uw.edu

ABSTRACT

While there are many existing tools for visualizing data, these tools have proven ineffective for analyzing the data collected by the Emerging Capacities of Mass Participation (emCOMP) Lab at The University of Washington. This research group, which explores the perpetuation of rumors on Twitter during crisis came to us with the need for a way of visualizing not only affirmations and denials of rumors over time, but also who participates in rumoring at a higher level (which accounts are responsible for higher volumes of re-tweets) while being able to understand potential reach and spread over time. With their needs in mind, we developed The Spaghetti Model. This paper discusses our design methods and implementation of the model.

Author Keywords

Social computing; social medial; rumoring; information diffusion; crisis informatics; reputation management

ACM Classification Keywords

H.5.3 Information Interfaces & Presentation: Groups & Organization Interfaces: Collaborative computing, Computer-supported cooperative work

INTRODUCTION

The Emerging Capacities of Mass Participation (emCOMP) Lab at University of Washington focuses on the behavior of the masses on social media during crisis events. One recent point of interest for their research has been the identification of user accounts that play a large role in the propagation and correction of crisis-related rumors on Twitter. In order to acquire the data pertaining to this interest, the researchers in the emCOMP Lab use a mixed methodology that involves MongoDB search strings to get relevant data and then Python scripts that have to be manually updated for each account. Though this method gives them the data they need to make their analysis, it is a manual, time-consuming process with a high potential for error. Our solution aims to eliminate the need for the above method by connecting the data directly to a visualization that reveals the information of interest.

BACKGROUND

Data Domain

Collected with the Streaming Twitter API, the data used by the emCOMP lab is not initially (structurally) different from standard Twitter data format. After collection however, additional data is added. This data consists of codes that are manually assigned to each individual tweet by a team of trained coders. Codes are assigned at two different levels. The first level consists of five options of which only one can be chosen: affirm, deny, neutral uncodable and unrelated. The second offers three options, all of which can be selected for one tweet: implicit, uncertainty and ambiguity. For the purposes of this model, the codes of interest are affirm and deny only

Though emCOMP has collections on many events and their resulting rumors, we focused on one rumor, which stemmed from what is referred to as the Sydney Siege Event.

Existing Methods Used

In order to determine which accounts are responsible for perpetuating rumors, members of the research group use the following process:

First, members of the Data Science sub-group graph affirmation and denial volume over time in Tableau (Figure 1)

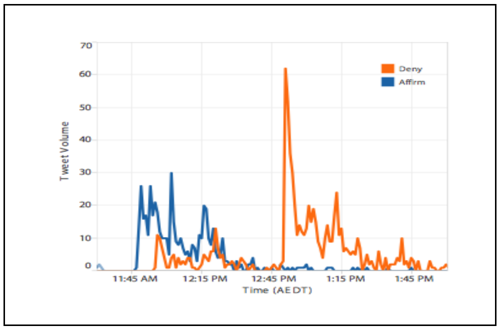


Figure 1. Tweet Volume over Time by Code: Affirm, Deny Lakemba Raids Rumors

Second, they investigate peaks in the graph with high tweets per minute (TPM) and locate user accounts that are being heavily re-tweeted at that time. These accounts are identified using the following MongoDB search string:

```
db.collection.find({created_ts : {
"$gte" : ISODate("yyy-mm-ddThh:mm:ss"),
"$lt" : ISODate("yyy-mm-ddThh:mm:ss")
}},{"_id" : 0,"retweeted_status.user.
screen_name" : 1}).sort({"retweeted_
status.user.screen_name" : 1}
```

This string works with the standardized retweet information embedded in tweet metadata. It works by pulling the user name of any account that has existed in the retweet field. That is, if a tweet is a retweet of a retweet, the retweeted_status.user.screen_name” field would include both the username of first account that was retweeted and the user name of the second account.

Once this information has been pulled, researchers manually look for the repeated presence of a given username at the time specified. If an account is repeatedly referenced in the field it is considered to be of interest.

Visualizing Accounts of Interest

After a satisfactory amount of retweet volume has been allocated to a group of individual accounts of, their individual volumes are graphed on a time-series plot with each account’s retweet volume represented by area, encoded with values dependent on their code and measured against the total tweet volume over time (Figures 2 & 3)

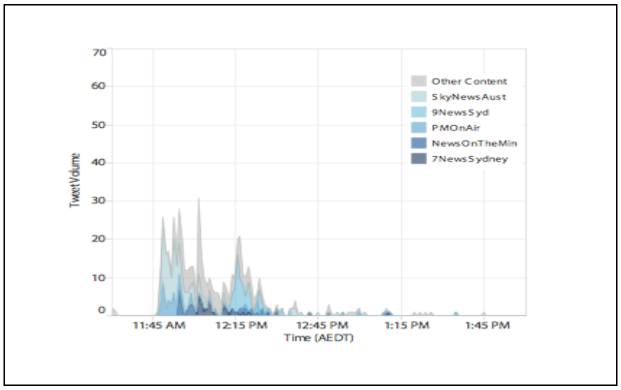


Figure 2. Affirm Volume over Time by Retweeted Account Lakemba Raid Rumor

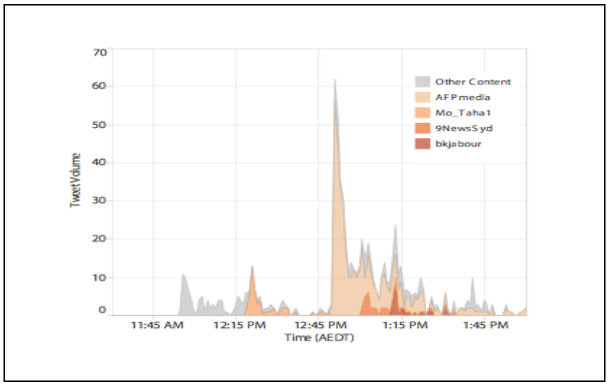
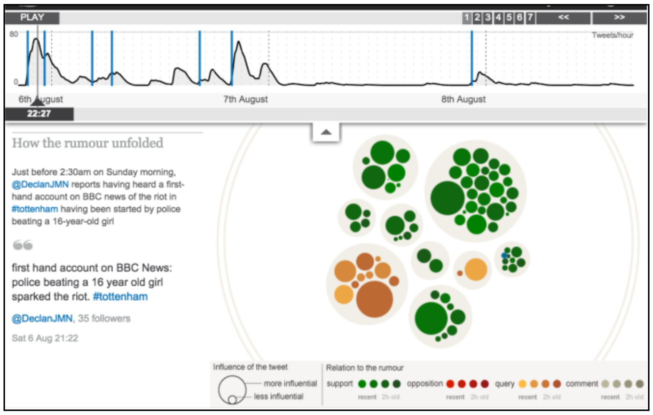


Figure 3. Deny Volume over Time by Retweeted Account Lakemba Raid Rumor

RELATED WORK

The current landscape of visualizations that examine rumoring on Twitter is scarce and we found only one satisfactory example that addresses rumoring on Twitter and how information spreads during crisis events. Developed by the Guardian, this visualization (Figure 3) focuses on the spread of rumors and how the sentiment of participants changes over time and its prime directive is to “show the birth and death of rumours on Twitter” [1]—



How the rumour unfolded

Just before 2:30am on Sunday morning, @DeclanJMN reports having heard a first-hand account on BBC news of the riot in #tottenham having been started by police beating a 16-year-old girl

first hand account on BBC News: police beating a 16 year old girl sparked the riot. #tottenham

@DeclanJMN, 35 followers

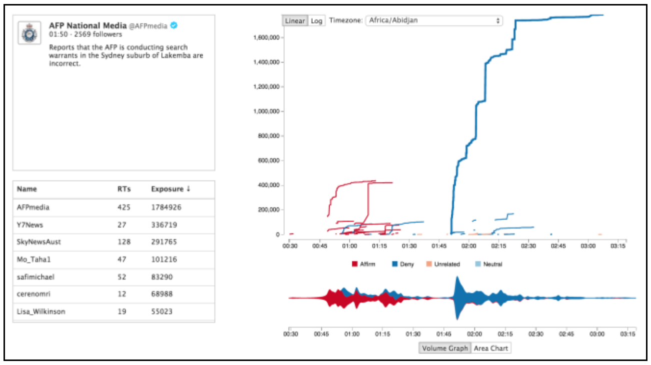
Sat 6 Aug 21:22

Figure 4. Behind the Rumours: how we built our Twitter riots interactive.

In this interactive example, tweet volume over time is displayed in the top panel, how the rumor unfolded is shown in the left panel, which updates at pivotal moments (marked with blue on the above time series) while the tweet of interest at the time is displayed below. In the main panel, which morphs as the animation plays and time passes tweets are clustered based on similarities in the text of the tweets. The influence of an author, which is determined by follower count is indicated by the size of a given circle and whether it supports, opposes, questions or comments on a rumor is indicated using color and value. The play function allows the story of rumoring to be told over time and the volume timeline allows users to see peak periods of conversation and if desired, focus only on a certain time.

DESIGN

The main goal of our design was to eliminate the emCOMP Lab’s use of MongoDB search strings to locate influential accounts by providing a tool that allows for visual exploration of that data. By connecting their data collection directly to our visualization and replacing their manual method with scripts, we were able to create a tool that does just that.



AFP National Media @AFPmedia
01:50 · 2569 followers

Reports that the AFP is conducting search warrants in the Sydney suburb of Lakemba are incorrect.

Name	RTs	Exposure ↓
AFPmedia	425	1784926
Y7News	27	336719
SkyNewsAust	128	291765
Mo_Taha1	47	101216
safimichael	52	83290
cerenomri	12	68988
Lisa_Wilkinson	19	55023

Figure 5. The Spaghetti Model

Features

The main panel of this visualization tool displays the retweet lifecycle of individual tweets over time. Each tweet is represented by a line, which is encoded using color to indicate its corresponding code. A tweet travels from left to right over time on the x-axis its popularity (an aggregate the follower counts of all accounts that have retweeted that tweet) is measured on the y-axis. For this panel, we included the option to change from views log (Figure 6) scaling and linear (Figure 7) scaling on the y-axis to allow for exploration of tweets that gained little or much traction.

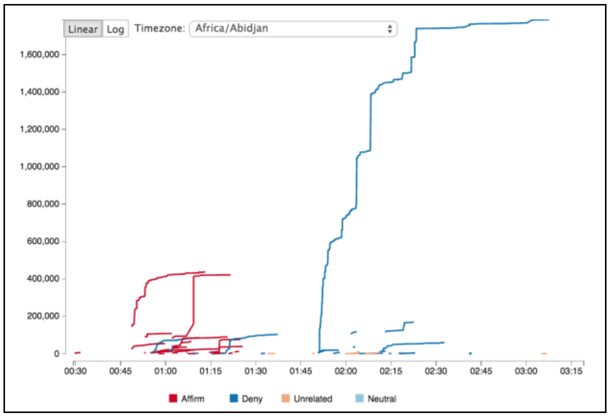


Figure 6. The Main Panel with Linear Scaling.

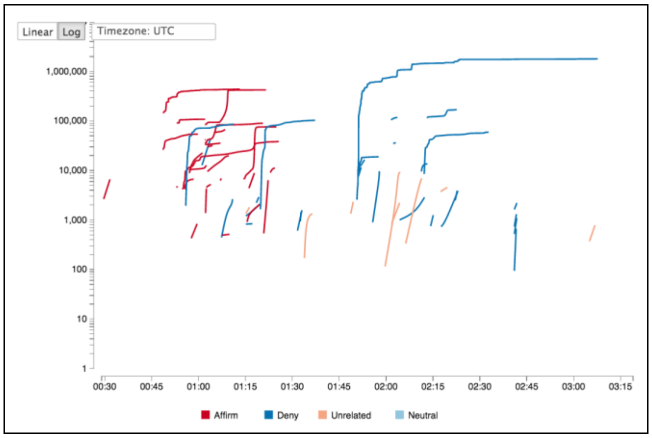
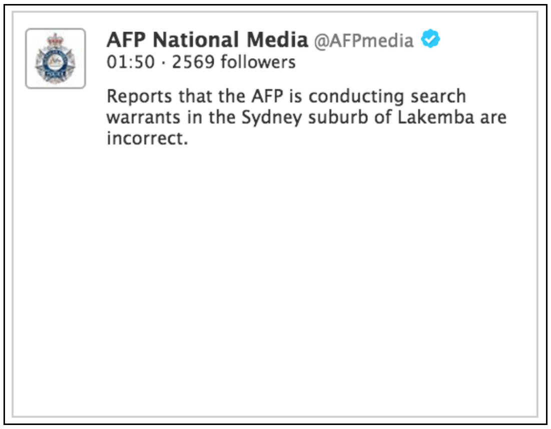


Figure 7. The Main Panel with Log Scaling.

In order to get information about an individual tweet, users can hover over a representative line, which will change in thickness. When a tweet is selected (clicked on) it has some permanence unsupported by hovering. When selected, the representative line width remains thicker than other tweets in the panel.

Additionally, information about the individuals who have perpetuated that tweet appears beneath the original tweet. This information includes their screen name and follower count (Figure 8). The list is per default sorted by retweet time, but it can be sorted by any of the attributes. On hovering one of the rows, the position of the respective retweet is marked with a dot on the representative line.



AFP National Media @AFPmedia
01:50 · 2569 followers

Reports that the AFP is conducting search warrants in the Sydney suburb of Lakemba are incorrect.

Figure 8. An Example of the Additional Information that Populates About Individual Accounts

The stream graph, located below the main panel allows users to explore specific periods of time. Using brushing, when a selection is made on this graph, that time period is reflected in the main panel. Two views are provided in order to allow for a better understanding of the dominant code during periods of low volume.

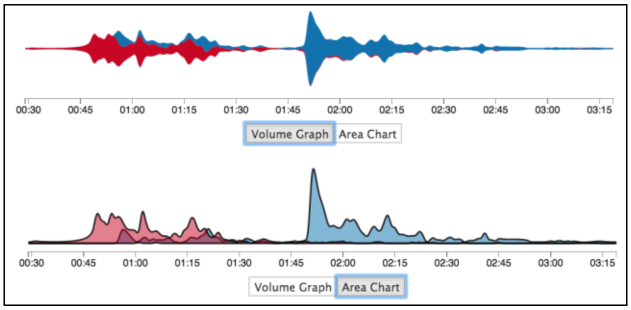
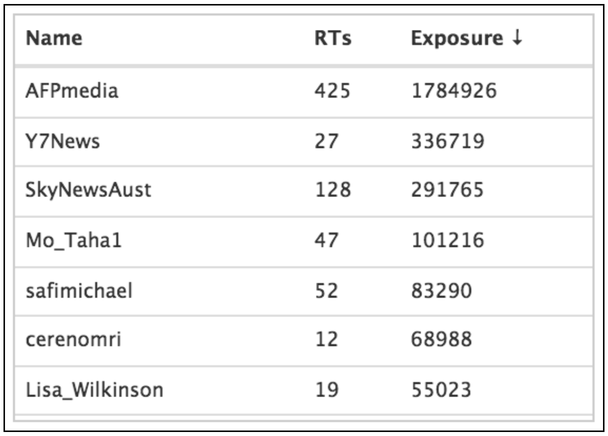


Figure 9. Both Views of the Stream Graph

The leaderboard, in the right-left corner of our dashboard, allows users to view an overall summary of the most influential accounts across the rumor.



Name	RTs	Exposure ↓
AFPmedia	425	1784926
Y7News	27	336719
SkyNewsAust	128	291765
Mo_Taha1	47	101216
safimichael	52	83290
cerenomri	12	68988
Lisa_Wilkinson	19	55023

Figure 10. The Leaderboard

Each account that emits an original tweet is assigned a score that represents the cumulative sum of the number of retweets inside the currently selected time period. The time zone of our visualization defaults to UTC but can be offset arbitrarily by selecting a time zone from a list. This feature enables the user to see the tweets in the local time zone of the main event or in a different time zone, as the rumor might have spread to different world regions. All the times displayed in the visualization will be displayed in the selected time zone.

Method

We found inspiration for this design in an exploration of the interactive capabilities of D3 (Figure11) made using randomly generated time series data. The bottom chart is used to navigate the main plot using the viewport, which allows users to choose a time period of interest and investigate that data further. This viewport is adjustable and can also be removed completely. The combination of charts paired with interaction allows users to see the relationship of the main chart to the overall data set.

The main panel is implemented as a D3 component, which facilitates redrawing after a viewport change. Highlighting the currently closest line on mousing over the panel is accomplished by overlaying a Voronoi tessellation of all the points of the line. This geometrical data structure assigns to every point in the space of the chart the closest point on one of the lines, which is then highlighted by referencing it. For both the linear and the logarithmic view, tessellations have to be computed, which are switched as needed.

The retweet list and the leaderboard make use of a table component written in D3. After trying out several libraries for drawing tables with exchangeable data, we found that none of those would fit our needs, as they were either bloated or hard to customize. Therefore, we decided on implementing our own. It features a fixed header. Clicking one of the header fields allows changing the sort order to the attribute of the respective column. Another click on the same header field toggles ascending / descending sort order.

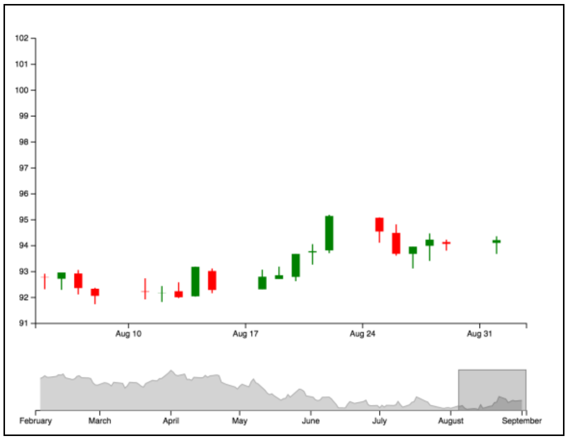


Figure 11. An Interactive Chart with D3

FUTURE WORK

We met with Kate Starbird multiple times throughout the development of this process, and her regular feedback helped in creating something that would be useful to her and her lab. However, there are some features that she requested that were not in the scope of this project timeline. We might implement these future enhancements to make the tool even more useful. In addition to some minor UI tweaks, some of these proposed additions are:

1. A tutorial for new users

Since this is a very specialized use case, we would like to provide some sort of tutorial for new users that would outline what this tool does and how to use it to make the onboarding process easier.

2. The ability to make this generalizable for any dataset

This would be a really useful feature because right now, our tool only shows one dataset. However, Kate's lab works with many datasets and making this tool generalizable would allow it to be used more widely. We've started implementing this but aren't finished with it yet.

3. A way to trim the dataset

Many datasets we've worked with have a lot of irrelevant tweets that would be better to just filter out. If we had a UI tool to allow researchers to crop a dataset, it would prevent them from having to mess around with code, which would save them time.

CONCLUSION

In partnership with Kate Starbird and the Emerging Capacities of Mass Participation Lab, we created an interactive visualization that effectively communicates the data (which accounts are responsible for higher volumes of re-tweets) of interest to emCOMP and achieved our goal to eliminate the inefficient mixed methodology previously used to identify this data.

ACKNOWLEDGMENTS

We thank Kate Starbird and the Emerging Capacities of Mass Participation Lab for their assistance and guidance.

REFERENCES

1. Behind the Rumours: how we built our Twitter riots interactive. Retrieved May 12, 2015 from https://about.twitter.com/company
2. Andy Aiken: Creating An Interactive Chart With D3. Retrieved May 12, 2015 from http://www.theguardian.com/news/datablog/2011/dec/08/twitter-riots-interactive
3. Cynthia Andrews, Elodie Fichet, Stella Yuwei Ding, Emma Spiro. 2015. Keeping Up with the Tweet-dashians: The Impact of ‘Official’ Accounts on Online Rumoring. Unpublished CSCW 2016.