TRAILS: A System for Monitoring the Propagation of Rumors On Twitter

S. Finn, P. T. Metaxas,* E. Mustafaraj, M. O'Keefe, L. Tang, S. Tang, L. Zeng
Computer Science Department
Wellesley College, Wellesley, MA 02481

ABSTRACT

Social media has become part of modern news reporting, whether it is being used by journalists to spread information and find sources, or as a medium by citizen reporters. The quest for prominence and recognition on websites like Twitter can sometimes eclipse accuracy and lead to the spread of false information. As a way to study and react to this trend, we introduce TRAILS, an interactive, web-based tool that allows users to investigate the origin and propagation characteristics of a rumor and its denial, if any, on Twitter. Propagation, timeline, retweet and co-retweeted network visualizations help users trace the spread of a story. While we envision that TRAILS would be valuable as a tool for individual use, in the initial stages we see it as a tool for amateur and professional journalists investigating recent and breaking stories. We are describing the system specifically for Twitter because it is easy through their APIs to collect the data. Given appropriate APIs we can build a system for Facebook and other social networks. Unfortunately, such APIs are not available at the time of this writing.

1. INTRODUCTION

The so-called "24 hour news cycle," has led to an increased sensationalism of news stories. Especially with the increase in cable news channels and online news media, the need to catch the attention of the public has led to faster and more hyped up reporting. Many compete to be the first to report a breaking story and present new and exclusive angles [5]. This trend has fed off social media and in turn empowered citizen journalists publishing and transmitting news through websites like Twitter and Facebook. Most of the time the information is true, but the desire to be first and receive more likes and retweets sometimes trumps accuracy and fact checking. It many cases, it may not matter much whether a rumor is true or false, but there are some cases that it matters greatly.

Consider the following scenario, that will serve as a running example in our description: Around noon on March 27, a reporter sees a tweet indicating that an airplane was spotted in the sea near the Canary Islands. For context, this happens just a few weeks after the disappearance of the Malaysian Airlines 370 flight on March 8, 2014, which captured the attention of people world wide.

Figure 1: A tweet spreading around 12 noon EST on March 27, 2014, reads (in Spanish) "Picture of the airplane in the sea these moments in Telde, Grand Canary Island".

Pressing the retweet button is very tempting in the situation, but spreading this information further should not be done automatically. It would be very helpful if the reporter can determine quickly a few facts about this story, including:

- **Originator**: Who posted the information first?
- **Burst**: When and how did the story break?
- **Timeline**: Is the story still spreading at the time of the inquiry?
- **Propagators**: Who has been retweeting and spreading the story, given the retweets often indicate agreement [4].
- **Negation**: Were there denials of the story competing for attention?
- **Main actors**: Who were the main actors in the propagation, according to the Twitter audience?

*Corresponding author.

In this paper, we present TRAILS, a new web-based tool for interactive exploration of Twitter information, which helps answer the above-listed questions. TRAILS retrieves relevant data from Twitter based on inputs from the tweet the user is investigating, and provides propagation, timeline and network based interactive visualizations.

TRAILS can be useful to reporters who discover some interesting information on Twitter, but are in the dark about its source and credibility. Knowing the answers to these questions can help decide how likely it is for the information to be true and whether it should be broadcasted further (with appropriate disclaimers).

1.1 Trails of Propagation

Credibility of information is strongly related to trust in the source. Before a savvy Twitter user retweets a tweet, she should feel reasonably confident in the validity of the information presented in that tweet, or else she might risk damaging her own reputation. This is true for aspiring citizen journalists, and even more important for professional journalists on social media. In fact, journalists care about informant credibility more than the average user [4]. TRAILS is an investigative and exploratory tool, to analyse the origin and spread of a rumor on Twitter. While it does not answer directly the question of its validity, it provides information that a critically thinking person can use to examine how a Twitter audience reacts to the spreading of the story. We currently envision TRAILS as a tool for journalists utilizing Twitter as a source of information, but in the future we want it to be useful to Twitter users with a working understanding of our visualizations.

TRAILS takes, as an input from the user, a single tweet with information she wishes to investigate, and allows the user to input keywords from that tweet to collect a set of related tweets. From that set of related tweets it provides visualizations to pinpoint the origin of the investigative tweet: where the information trail started and who initially reported it. In many cases this may be enough for the user, based on the reputation of the accounts which broke the story on Twitter, by weighing factors such as whether they are verified, if they have many followers, the age of the account, or studying their profile and recent tweets.

In cases of more dubious data, or for a more engaged Twitter user or journalists, TRAILS provides visualizations to trace not only the origin, but the spread of a story. It gives the user tools to answer interesting questions about the story: who started this story, and who popularized it? When did the story break and how did it spread? When was it most active, and what information and users were prominent at peak times? And what users were influential in the spread of the story, and who did users put their trust in when spreading the story? Propagation and Timeline visualizations give the user a meaningful way to browse the data, while network graphs give her an overview of influential users in the data.

2. RELEVANCY AND BURSTINESS

2.1 Tweet Relevancy Algorithm

The goal of TRAILS is to discover and visualize the rumor behind a tweet: to find where the information in the tweet originated, how it propagated and manifested over time and who had a hand in spreading it. The first step to creating a story is to collect the data which will form the story. TRAILS is focused on the spread of a rumor via Twitter, so the stories it creates will be made up of tweets which are *relevant* to the user's investigation. It is important that the definition of a "relevant" piece of data should be broad enough to capture interesting and important facets of the story, but limited so that the dataset is manageable for human consumption. TRAILS attempts this task by employing a user-controlled data filter based on keywords from the investigative tweet and the user's own knowledge.

The user may control the collection of data by selecting words or phrases (which we refer to as "keywords") from the text of the tweet they are investigating, or by manually inputing keywords absent from the text. TRAILS then takes the keywords chosen by the user to collect tweets via Twitter's Search API, with each keyword being used as a search query. The number of tweets collected can be controlled by the user as well. The Search API returns recent tweets containing words in the query in reverse chronological order (newer tweets are returned first), but is limited to tweets written or retweeted in the last 6-9 days. Because of this limitation, TRAILS is best suited to investigating recent and breaking stories.

From the data retrieved by the Search API, TRAILS automatically calculates which are the relevant tweets based on other inputs the user can control. The user can require a relevant tweet to contain some subset of keywords. The user defines keywords as being either "required", "optional" or "excluded." Required keywords must appear in a tweet for it to be relevant, and if the user choses multiple required keywords, she can select whether all or at least one must appear in a relevant tweet. The optional keywords are controlled by a threshold (also defined by the user): a relevant tweet must contain at least the number of optional keywords set by the threshold. The threshold can be set to 0, in which case the optional keywords have no effect on the relevancy algorithm. Finally, any tweet containing any of the excluded keywords will not be considered relevant (even though it may have the optional or required keywords). In addition to the keyword-based inputs, the user can define a time period to limit relevant tweets. Any tweets outside of the time period will not be considered relevant.

In many investigations the initial set of search keywords must be modified to define relevant tweets. TRAILS gives the user the option to redefine this set as many times as she wants until she has discovered the best set of relevant tweets to study.

2.2 Burstiness Algorithm

One of the purposes of TRAILS is to give users tools for investigating the origins of a story on Twitter: who broke the story and when. TRAILS automatically identifies the moment a story breaks on Twitter by computing the time interval in which relevant activity in the story increases significantly. For brevity we will skip the details of the mathematics of the computation here. We consider activity as the sum of the retweet counts because this correlates to how much attention a tweet received, and how many people have been exposed to it. Our algorithm does not identify the first tweet to post relevant information about the story, but the first tweet with the greater impact; if a first tweet received very little attention, then we don't consider it to have broken the story.

Once we have identified the initial burst in the data, we visualize these tweets in the "Propagation Graph" described in Section 3.3.1, to allow the user to study them in more detail and answer questions about how the story originated and how information propagated when the story broke.

2.3 Negation Algorithm

Another question TRAILS allows the user to investigate is whether there a denial or negation of their story circulating, and how it has

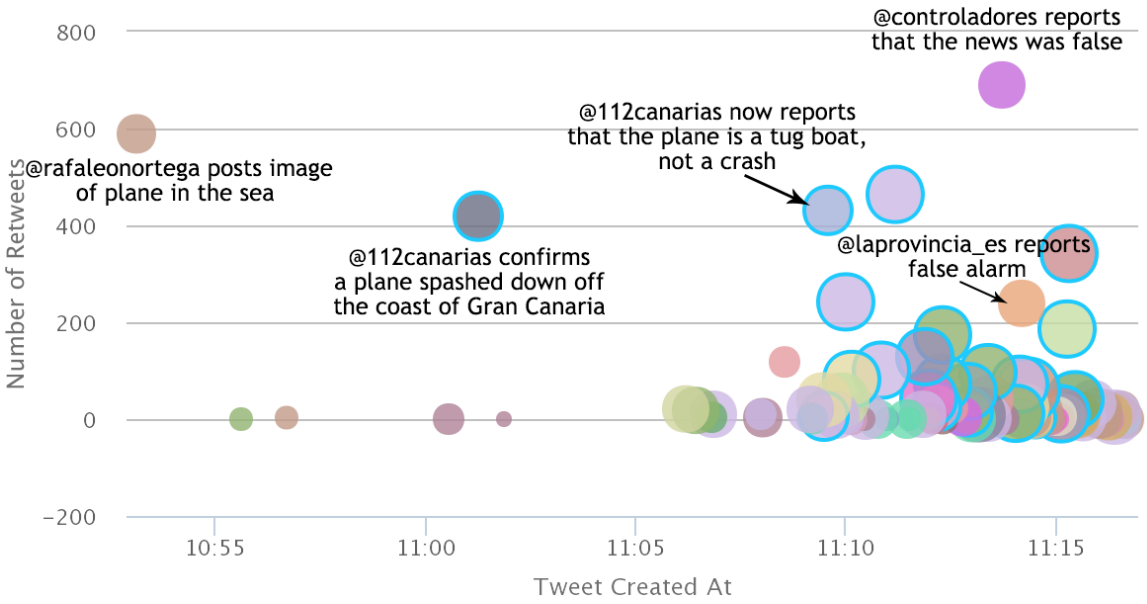

Figure 2: The Tweet Propagation Graph from the Plane in the Sea story shows the first 100 tweets after the burst in propagation.

spread. We do this by calculating "negation" tweets. Currently we employ a simple algorithm which identifies tweets with a small set of keywords that indicate negation, doubt, or denial, such as "hoax," "fake," and "untrue." In addition, the user can add to or remove these keywords on a story by story basis. These negation tweets can then be visualized in the Timeline graph (described in Section 3.3.2).

3. ARCHITECTURE OF THE SYSTEM

TRAILS is designed to provide users with an interface to investigate the origin and propagation of a tweet. This section will describe the different tools and visualizations TRAILS gives users.

3.1 Investigative Tweet

TRAILS is structured around the investigation of a single tweet, which is the first input the user provides (via the url of the tweet). Throughout this paper we will reference the Plane in the Sea story, which was investigated following a tweet by @rafaleonortega on March 27th, 2014, reporting that there was a plane in the sea off the coast of Telde in the Canary Islands (see Figure 1). The tweet includes a picture of what looks like a plane in the water.

After selecting the investigative tweet, TRAILS provides the Keyword Selection interface, to allow the user to highlight words and phrases from the tweet as keywords, or enter them manually. It also *suggests keywords* by querying Twitter for the 100 most recent tweets containing keywords the user has already selected and displaying terms that appear frequently in the results. For @rafaleonortega's tweet, we chose the following keywords: "gran canaria", "imagen", "airplane", and "telde".

3.2 Refining Relevant Tweets

The algorithm for searching for, collecting and determining relevant tweets is descibed in section 2.1. TRAILS provides an interface for the user to modify the inputs which go into the determining the relevant tweets as many times as she likes in order to select the best set of relevant tweets (described in section 2.1). For the Plane in the Sea story, we added "avión" as the only required keyword, with none of the other search terms needing to be present.

3.3 Story Interface

3.3.1 Tweet Propagation Graph

The first tool we present to the user is the Propagation Graph: a novel visualization which uses data from when the story broke on Twitter to show who originated the story, and highlight influential and independent content creators. The burstiness algorithm (described in section 2.2) is used to identify the time when the story breaks, and the propagation graph shows the first hundred tweets in the breaking interval (see Figure 2). A data point in the Propagation Graph represents a single tweet, and is plotted in several dimensions: the x-axis shows *time*; the y-axis shows the *number of retweets* retrieved; and the size of a point represents the *number of followers* the tweeter has (scaled logarithmically). Tweets written by *verified accounts* are marked by a bright blue border. We claim that these are key elements in gauging the visibility of the tweet, as well as the degree of credibility other users will assign to the tweet and amount of trust in the user as a source of information. Since we are trying to track the flow of a story, time is a natural factor to observe.

Another facet of the graph is the color of the data points: tweets with *similar language* will have the same color, in an attempt to visualize independence of data. Many tweets without variation in their wording may be a reason for suspicion. The tweet text might be the headline of an article with short commentary, a single tweet being copied and modified, or even a single person spamming the same information and varying the language very little. More variation in color likely means there are likely multiple sources twitting about the story: several different articles, or many individuals using different phrasing to talk about the same subject.

The web interface allows users to view the tweets represented by data points on the graph by hovering over or clicking on the data points. Studying the Propagation Graph (fig. 2), we discover some

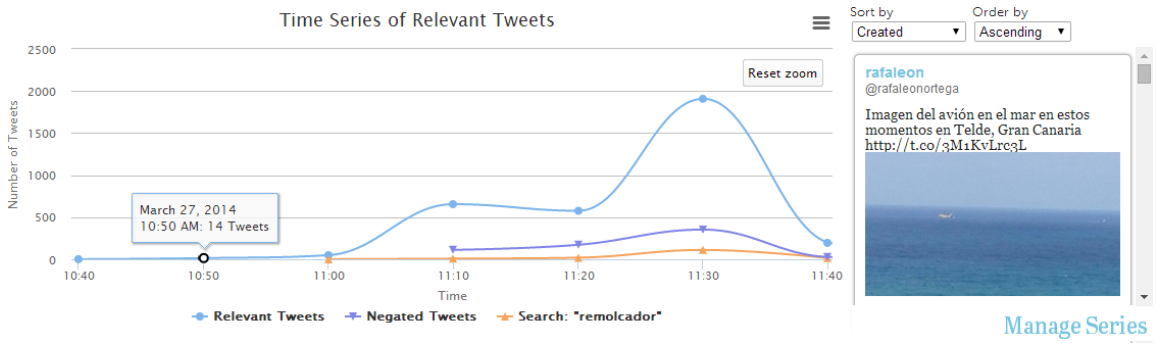

Figure 3: The Timeline visualization from the "Plane in the Sea" story. Selecting a data point brings up a pane with all the tweets sent during this 10-minute interval. Three series are shown in this graph: all the relevant tweets, the negating tweets, and those the user chose to search for containing a particular keyword: remolcador (tug boat). It appears here that the negating tweets have succeeded in affecting the propagation of the rumor.

facts about how the "Plane in the Sea" story developed. Looking at the graph as a whole, we can see that the tweets, spread over only 20 minutes, are varied in content (many different colors) and number of retweets, and the users who have written them also vary in the number of followers they have. There is also a certain number of verified users, mostly news organizations.

In this case study, the breaking tweet appears as the first one in the graph. It happens to be the investigative tweet, reporting at 10:53 am EST that there is a plane in the sea near Telde in the Canary Islands, with a blurry picture of what appears to be an airplane in the water. Information about the originator, the user who started the rumor by tweeting the picture of the airplane, is provided as well: @rafaleonortega's describes himself as a sports reporter, and has a small number of followers. @rafaleonortega is not perhaps the most credible source for breaking news, but his tweet has almost 600 retweets, so his message has been fairly well propagated, likely due to the accompanying image.

The next few tweets have similar messages, talking about a plane crashing in the sea, including a tweet at 11:01 am from @112canarias, a verified account tweeting about emergency information in the Canary Islands (112 is similar to 911 in the US). This tweet confirms that a plane splashed down off the coast, though they do not know the number of passengers. However, less than ten minutes later, at 11:09 am, @112canarias tweets again, now reporting that what was mistaken for a plane is actually a tug boat; at the same time, other verified accounts continue to report that a plane has crashed in the sea. Two tweets from unverified accounts (@controladores at 11:13 am with 690 retweets, and @laprovincia_es at 11:14 am with 239 retweets) also report that the plane crash is false, while more accounts continue to report about the crash.

3.3.2 Timeline Visualization

The Propagation visualization gives a detailed look into a specific interval of time: when the story broke; the Timeline visualization gives an overview of the whole story. An example is shown in Figure 3. The user can selectively browse the data in the timeline visualization without being overwhelmed by thousands of tweets. Selecting a point on the graph will display the tweets written in that 10-minute interval to the right of the data point. They can be sorted in ascending or descending order by the number of retweets received, the time they were created or whether they are original tweets or retweets. The negation tweets (discussed in Section 2.3) are also displayed as a series in this graph, to show when tweets denying the story began to spread.

Although the propagation graph shows @rafaleonortega breaking the story, the time series shows a tweet written four minutes earlier by @SilviaLuzHernnd claiming there was a plane crash, and mentioning two emergency information accounts, @112canarias and @Infoemergencias.

The story begins to pick up in popularity after 11:10 am, with over 500 tweets in a ten minute period. The story peaks in popularity at 11:30 am, with almost 2,000 tweets in ten minutes. The first mention of a tug boat can be seen at 11:09 am, preceeding the negating information which claims the story of the crash is false. Although the number of tweets negating the rumor never equals the number claiming it to be true, as the number of negating tweets increases, the number supporting it begins to decrease.

3.3.3 Retweet and Co-Retweeted Networks

The first two visualizations focus on the propagation of content over time. TRAILS also allows users to investigate the relationship of networks of users retweeting the story. The next two visualizations, the Retweet Network and the Co-Retweeted Network, help to answer questions about the main actors who were spreading information.

The two networks are created using Gephi [1], where each node represents a user. In the retweet network (Fig. 4), an edge between two nodes represents one user having retweeted the other. The edges curve clockwise from the retweeting user to the retweeted user. In the co-retweeted network on the other hand (Fig. 5), edges represent two users having both been retweeted by a third user (see [2] for more discussion on the co-retweeted algorithm). Larger and darker nodes have a higher degree, and nodes are colored based on the modularity algorithm which groups nodes which are closely connected in the network [2].

Figure 4 shows the Retweet Network Visualization of the Plane in the Sea case study. On the left is the retweet network graph, while on the right is an information panel which describes the node which the user has clicked on (in this case, @laprovincia_es): it shows information supplied by the user on their Twitter profile, as well as information about their tweets in the relevant tweet dataset.

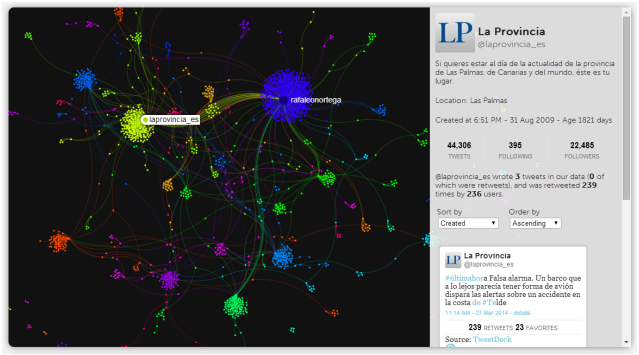

Figure 4: The retweet visualization from the Plane in the Sea story shows the retweeting activity in the network of users.

The main actors, users who received the most retweets, appear largest and most prominent in this graph. @rafaleonortega appears as the most retweeted node in the graph, by 554 different users in the dataset. @laprovincia_es is also highly visible, with 236 users retweeting its tweets. These two users were highlighted before in the Propagation visualization (Figure 2), where @rafaleonortega has spread the rumor of the plane crash, and @laprovincia_es tweeted denying the rumor. Note that the clusters of accounts retweeting each of these two accounts are mostly not overlapping, indicating that each group has heard either the original story or its denial. However, there is a smaller group of accounts retweeting both @rafaleonortega and @laprovincia_es. These are the users who propagated the initial false information and then propagated its correction. Clicking on these users on the web interface reveals that they have almost all retweeted first @rafaleonortega and then, minutes afterwards, retweeted @laprovincia_es' denial. This lends credibility to @laprovincia_es' information: even though less users have retweeted him, his information that the crash was a false story was conclusive for many users.

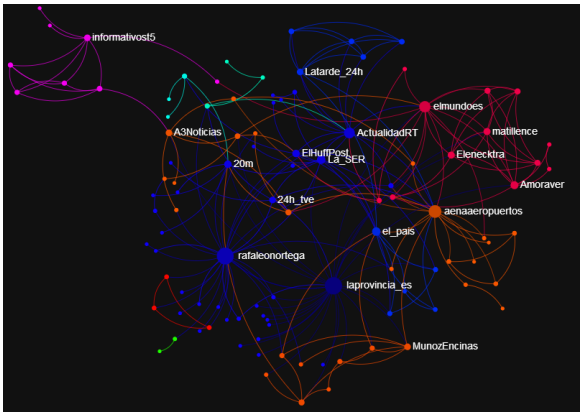

Figure 5: The co-retweeted network from the Plane in the Sea story shows the main actors, according to the users' retweets.

The co-retweeted network shown in figure 5 highlights the main actors from the retweet network, by connecting accounts based on mutual retweeting users. That is, if User A and User B in the co-retweeted network and connected by an edge, it means at least one other user has retweeted User A and retweeted User B. @rafaleonortega and @laprovincia_es are connected in the co-retweeted network because of the users who retweeted both of them. Connections indicate related content: in this case the relationship is that the content created by @laprovincia_es is a response and contradiction of information from @rafaleonortega.

4. CONCLUSION AND FUTURE WORK

TRAILS was designed and implemented with the goal to provide a vital service to users who want to engage with Twitter as a source of reliable information, either for their own consumption, or as a source for journalism, both professional and amateur.

TRAILS makes it easy to investigate a suspicious story. By inputting a single tweet into the system, and selecting keywords relevant to the story being investigated, the system will gather a dataset of tweets through which the user can trace the story origin. The Tweet Propagation visualization focuses on the moment the story first broke on Twitter, while the Timeline Visualization on how it spread. Both of these allow users to meaningfully and easily sort through hundreds to thousands of tweets, and highlight both tweets and periods of time that are most interesting to the story. The two network visualizations, a Retweet and a Co-Retweeted network, allow users to study accounts on Twitter who were both influential propagators of information, and sources who other users put trust in. See `http://bit.ly/twittertrails`

The most pressing area of future study for TRAILS is to design and implement a user evaluation of the tool, to further improve its functionality and usefulness. We plan to draw inspiration from Kang et. al [3], who outline a method for in depth evaluation of visual analytics systems. They log and analyse user activity not only using the system to be evaluated, but also with more low tech approaches to solving the same problem.

We also plan on evaluating and improving our algorithms to detect when a story breaks on Twitter, and filtering for relevant tweets. We also hope to pursue more methods of customizing TRAILS for users, in ways specific to the story they are investigating. This may include creating more visualizations, which the users can pick and chose from as is appropriate to their investigation, and creating more meaningful ways in which these visualizations can interact with each other.

5. ACKNOWLEDGEMENTS

This research was partially supported by NSF grant CNS-1117693 and by the Wellesley Science Trustees Fund.

6. REFERENCES

[1] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. In *2009*.

[2] S. Finn, E. Mustafaraj, and P. Metaxas. The co-retweeted network and its applications for measuring the perceived political polarization. In *WEBIST 2014*, 2014.

[3] Y.-a. Kang, C. Görg, and J. Stasko. Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In *IEEE VAST Symposium*, 2009.

[4] P. Metaxas, E. Mustafaraj, K. Wong, L. Zeng, M. O'Keefe, and S. Finn. Do retweets indicate interest, trust, agreement? Computation and Journalism Symposium, NYC, NY, 2014.

[5] S. Rosenbaum. Too much news? *Huff. Post*, Jan 2013.