Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing

Kate Starbird, University of Washington, kstarbi@uw.edu
Jim Maddock, University of Washington, maddock@uw.edu
Mania Orand, University of Washington, orand@uw.edu
Peg Achterman, Northwest University, peg.achterman@northwestu.edu
Robert M. Mason, University of Washington, rmmason@uw.edu

Abstract
The Boston Marathon bombing story unfolded on every possible carrier of information available in the spring of 2013, including Twitter. As information spread, it was filled with rumors (unsubstantiated information), and many of these rumors contained misinformation. Earlier studies have suggested that crowdsourced information flows can correct misinformation, and our research investigates this proposition. This exploratory research examines three rumors, later demonstrated to be false, that circulated on Twitter in the aftermath of the bombings. Our findings suggest that corrections to the misinformation emerge but are muted compared with the propagation of the misinformation. The similarities and differences we observe in the patterns of the misinformation and corrections contained within the stream over the days that followed the attacks suggest directions for possible research strategies to automatically detect misinformation.

Keywords: Crisis informatics; Twitter; microblogging, misinformation; information diffusion; rumors; crowdsourcing

Citation: Editor will add citation with page numbers in proceedings and DOI.

Acknowledgements: [Click here to enter acknowledgements]
Research Data: In case you want to publish research data please contact the editor.
Contact: Editor will add e-mail address.

1 Introduction

Social media use is becoming an established feature of crisis events. Affected people are turning to these sites to seek information (Palen & Liu, 2007), and emergency responders have begun to incorporate them into communications strategies (Latonero & Shklovski, 2011; Hughes & Palen, 2012). Not surprisingly, one concern among responders and other officials is the rise of misinformation on social media. In recent crises, both purposeful misinformation, introduced by someone who knew it to be false, and accidental misinformation, often caused by lost context, have spread through social media spaces and occasionally from there out into more established media (Hill, 2012; Herrman, 2012).

In a study on Twitter use after the 2010 Chile earthquake, Mendoza et al. (2010) claimed that aggregate crowd behavior can be used to detect false rumors. They found that the crowd attacks rumors and suggested the possibility of building tools to leverage this crowd activity to identify misinformation. However, currently there are no such tools, and the notion of the self-correcting crowd may be overly optimistic. After Hurricane Sandy, a blogger claimed to have witnessed the "savage correction" by the crowd of false information spread by an aptly named Twitter user, @comfortablysmug (Hermann, 2012), yet many were guilty of retweeting this and other misinformation during the tense moments when Sandy came ashore (Hill, 2012).

This research, which focuses on the use of Twitter after the 2013 Boston Marathon bombings, seeks to understand how misinformation propagates on social media and explore the potential of the crowd to self-correct. We seek to better understand how this correction functions and how it varies across different types of rumors. Our larger goal is to inform solutions for detecting and counteracting misinformation using the social media crowd.

2 Background

2.1 **The Event: 2013 Boston Marathon Bombing**

At 2:49 pm EDT on April 15, 2013, two explosions near the finish line of the Boston Marathon killed three people and injured 264 others (Kotz, 2013). Three days later, on April 18 at 5:10pm EDT, the Federal Bureau of Investigation (FBI) released photographs and surveillance video of two suspects, enlisting the public's help to identify them. This triggered a wave of speculation online, where members of the public were already working to identify the bombers from photos of the scene (Madrigal, 2013a). Shortly after the photo release and a subsequent related shooting on the MIT campus, a manhunt resulted in the death of one suspect and the escape of the other. Following the shoot out, at 6:45 AM on April 19, the suspects were identified as brothers Tamerlan and Dzhokhar Tsarnaev (FBI, 2013). Dzhokhar, the surviving bother and "Suspect #2" from the FBI's images, was found and arrested on April 19 at 9pm EDT.

2.2 **Social Media Use during Crisis Events**

Researchers in the emerging field of crisis informatics have identified different public uses of social media during crises: to share information (e.g. Palen & Liu, 2007; Palen et al., 2010; Qu et al., 2011), to participate in collaborative sense-making (Heverin & Zach, 2012), and to contribute to response efforts through digital volunteerism (Starbird & Palen, 2011). Social media are a potentially valuable resource for both affected people and emergency responders (Palen et al., 2010). Twitter in particular has been shown to break high-profiles stories before legacy news media (Petrovic et al., 2013). This research focuses on misinformation (false rumors) shared through Twitter in the aftermath of the Boston Marathon bombing on April 15, 2013

2.3 **Misinformation on Twitter**

Misinformation on social media represents a challenge for those seeking to use it during crises. This concern has been voiced in the media (Hill, 2012) and by emergency responders who are reluctant to depend on it for response operations (Hughes & Palen, 2012). A few emergency managers who were early adopters of social media note that identifying and counteracting rumors and misinformation are important aspects of their social media use (Latonero & Shklovski, 2011; Hughes & Palen, 2012).

Mendoza et al. (2010) found that Twitter users question rumors, and that false rumors are more often questioned than rumors that turn out to be true. They theorized that this crowd activity could be used to identify misinformation.

2.4 **Diffusion of Information on Twitter**

An important aspect of the misinformation problem on social media relates to the diffusion of information. On Twitter, the retweet (RT @username) functions as a forwarding mechanism. During crisis events, a large percentage of tweets are retweets, which spread very quickly. Kwak et al. (2010) reported 50% of retweets occur in the first hour after a tweet is shared and 75% within the first day. As this information diffuses, it loses connection to its original author, time, and the context it in which it was shared, an effect that complicates verification.

3 Method

3.1 **Data Collection**

We collected data using the Twitter Streaming API, filtering on the terms: boston, bomb, explosion, marathon, and blast. The collection began April 15 at 5:25pm EDT and ended April 21 at 5:09pm EDT. During high volume time periods, the collection was rate-limited at 50 tweets per second. Additionally, the collection went down completely (Figure 1, black rectangle) and experienced two windows of repeated short outages (Figure 1, grey rectangles).

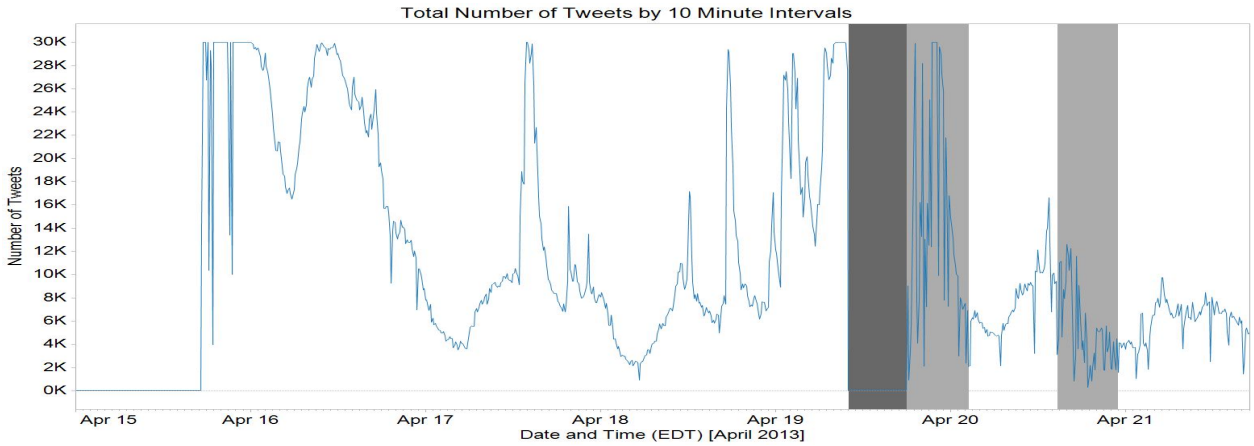

Figure 1. Tweet Volume Over Time

Our data set contains ~10.6 million (10.6M) tweets, sent by 4.7M distinct authors. 42% of tweets have URL links and 56% are retweets using either of the two most popular conventions (RT @ or via @). Figure 1 shows the volume of tweets per minute.

3.2 Exploratory Analysis: Identifying Rumors

Our exploratory analysis identified salient themes and anomalies in the data. First, we enumerated the 100-most prevalent hashtags in the data and created a network graph of relationships between them (Figure 2). Each node represents a popular hashtag and is sized by the log number of times in appears in the set. Each edge connects two hashtags that appear in the same tweet and is sized by the log number of times they co-occur.

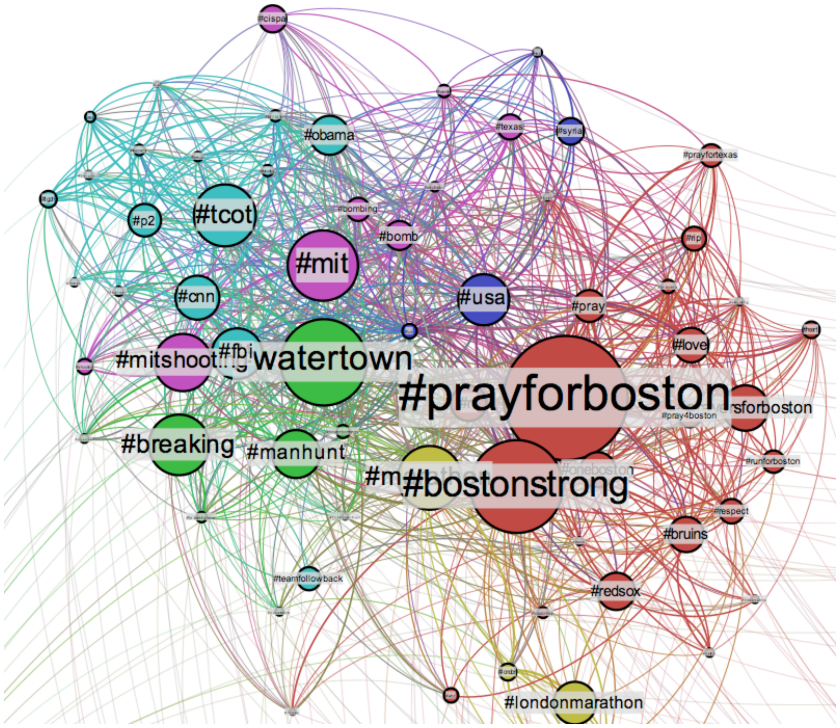

Figure 2. Network Graph of Co-Occurring Hashtags in Boston Marathon Tweets
*The #boston hashtag was dropped from this graph, because it connected with every other tag.

Next, we examined tweets that contained specific hashtags to understand their context. A salient theme was the presence of several rumors. For example, tweets with #prayforboston contained a highly tweeted (false) rumor about a young girl killed while running the marathon. And an investigation of a

politically themed section of the graph (in light blue at the top, left corner) revealed an interesting hashtag, #falseflag—positioned between #tcot (which stands for "top conservatives on Twitter) and #obama—that accompanied rumors claiming U.S. government involvement in the bombings.

Through this process, we created a list of rumors grounded in the Twitter data set. We chose three false rumors and proceeded to do a systematic analysis of tweets that referenced them.

3.3 **Analysis: Manual Coding of Tweets**

We selected search terms that resulted in samples that balanced comprehensive and low noise to identify tweets related to each rumor. Then, following the method outlined by Mendoza et al. (2010), we coded each distinct tweet within each rumor subset. We used an iterative, grounded approach to develop the coding scheme, eventually settling on three categories: *misinformation*, *correction*, and *other* (which encompassed *unrelated* and *unclear*). Our categories align well with the categories used by Mendoza et al. (2010): affirm, deny, other.

4 Findings

4.1 **Rumor 1: A Girl Killed While Running in the Marathon**

The most blatant false Twitter rumor focused on a photo of an eight year-old girl running in a race, accompanied by the claim that she died in the Boston attack. The rumor began just hours after the bombing. Its history on Twitter reveals that at approximately 6:30pm EDT, @NBCNews announced that an eight year old *spectator* had been killed in the bombings. About 45 minutes later, a Twitter user sent a message ascribing a female gender to the victim with the assumption that she was a competitor:

> **@TylerJWalter** (April 15, 2013 7:15pm): An eight year old girl who was doing an amazing thing running a marathon, was killed. I can't stand our world anymore

Four minutes later, another user added a fake picture and purposefully spread the false rumor:

> **@_Nathansnicely** (April 15, 2013 7:21pm): The 8 year old girl that sadly died in the Boston bombings while running the marathon. Rest in peace beautiful x http://t.co/mMOi6clz21

This original rumor was retweeted 33 times in our set, but it soon began to spread in many different forms, from different authors. We identified a set of 93,353 tweets (and 3275 distinct tweets) that contained both "girl" and "running." After coding each distinct tweet and applying those codes to the larger set, we found 92,785 tweets to be related to the rumor. 90,668 of these tweets were coded as misinformation and 2046 tweets were corrections, resulting in a misinformation to correction ratio of over 44:1. This finding contrasts starkly with Mendoza et al.'s (2010) study, which found about a 1:1 ration between the two.

Significantly, peak correction did occur roughly within the same hour interval as peak misinformation, suggesting reactionary community response. Examining the volume at log scale reveals the volumes of correction and misinformation to rise and fall in tandem much of the time, though the correction often lags behind the misinformation. Perhaps the most troublesome aspect of the graph shows misinformation to be more persistent, continuing to propagate at low volumes *after* corrections have faded away.

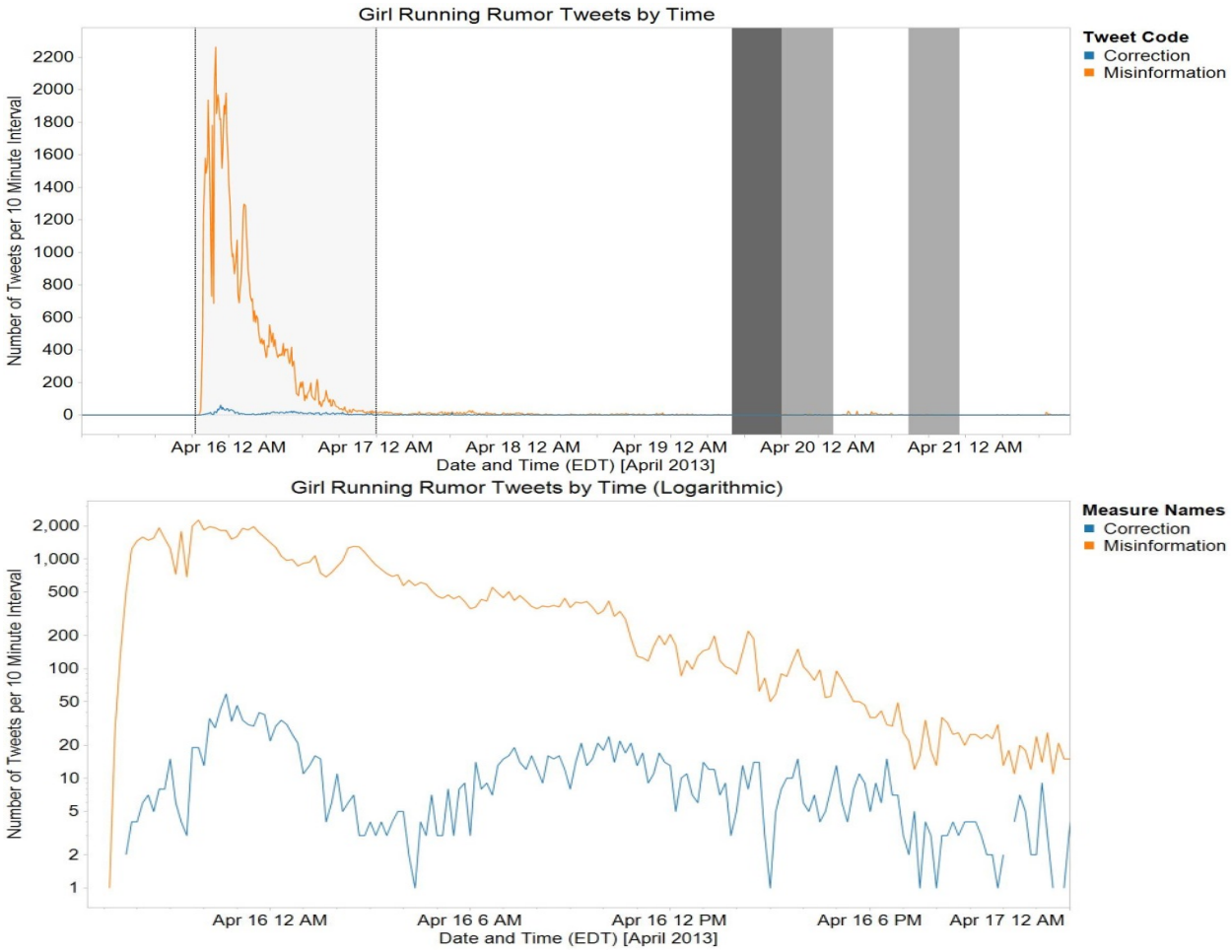

Figure 3. Girl Running Rumor, Tweets per 10 Minutes

*Light grey rectangle in top image highlights window of focus for bottom image

4.2 Rumor 2: False Flag — Navy Seals or Craft Security or Blackwater Agents as Perpetrators

After the FBI released images of dark, exploded backpacks, users of social media spaces like 4chan and Reddit began to collect and analyze images of suspicious individuals wearing backpacks at the scene. One set of images included two men standing together wearing the same clothes (caps, pants, and boots), carrying heavy dark backpacks of the style shown in the FBI photos. Some speculated these individuals might have been involved in a drill or in the actual attack (Watson, 2013). An emblem on one of their hats suggested to some that the men were affiliated with U.S. military special operations, specifically Navy SEALs. Later, speculation shifted to claims that they were agents of Blackwater or Craft International, both private military/security firms. Each explanation supported a larger claim—that the bombings had been a “false flag” attack, either staged or actually carried out by the U.S. Government (Watson, 2013).

For Rumor 2, we coded tweets that contained either “navy seal” or “blackwater” or “black ops” or (“craft” and “security”).

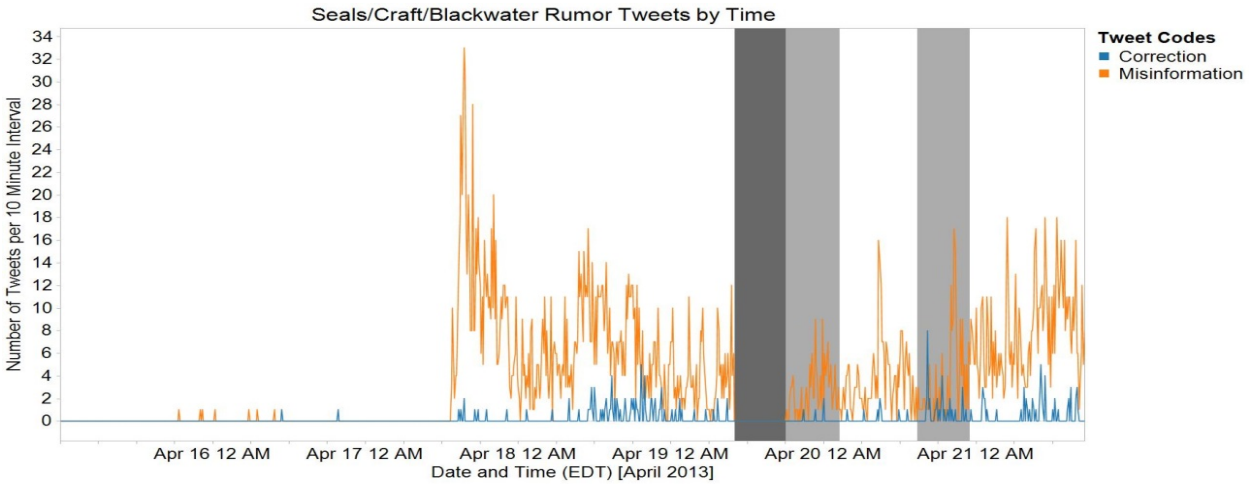

Figure 4. False Flag Rumor, Tweets per 10 Minutes

Rumor 2 had far fewer tweets than Rumor 1, only 4525 total. 3793 of these were misinformation, 212 were corrections, and 520 were coded as other, most of those being unrelated.

The diffusion of Rumor 2 appears to progress somewhat differently from Rumor 1, peaking once at the beginning but then persisting and eventually gaining steam again at the end of the collection period. However, some aspects of the relationship between misinformation and corrections are consistent with Rumor 1—volumes of both often rise and fall together, though there is often a lag between the spike in misinformation and the resulting rise in corrections. Again the ratio of misinformation to correction was high (18:1), though significantly smaller than for Rumor 1 (Chi-square with Yates correction, $p<0.0001$).

Because of the context of this rumor, it is unlikely that the corrections themselves had much effect in stemming the flow of misinformation. Even after the shoot-out, capture and identification of the Tsarnaev brothers, the rumor returned. Examining the content of correction tweets suggests that those who sent them were part of a separate conversation, criticizing the speculation but not interacting with those participating in it.

4.3 Rumor 3: Digital Vigilantes — The Crowd Misidentifies Sunil Tripathi as a Bomber

Sunil Tripathi was a 22-year-old Brown University student who went missing on March 16. In the weeks before the bombing, his family was actively searching for him, using social media and formal media outlets to raise awareness (Bidgood, 2013).

Within a few hours of the FBI releasing the grainy surveillance photographs of the bombing suspects, a few different Twitter users claimed that Tripathi looked like Suspect #2. In the most notable example, a former high school classmate of Tripathi's posted a tweet noting the resemblance (April 18 at 7:39pm). Shortly thereafter, a Reddit thread became focused around speculation of a connection between Tripathi and Suspect #2 (Reddit, 2013). The rumor continued into the early morning hours as the Watertown shootout was occurring. The following tweets fueled the spread:

> **@ghughesca** (April 19, 2:43pm): BPD has identified the names: Suspect 1: Mike Mulugeta. Suspect 2: Sunil Tripathi.
>
> **@KallMeKG** (April 19, 2:50pm): BPD scanner has identified the names : Suspect 1: Mike Mulugeta Suspect 2: Sunil Tripathi. #Boston #MIT

Neither statement was true, but within minutes many Twitterers, including some traditional media outlets, retweeted this misinformation (Madrigal, 2013b). Just before 6:40am, the FBI and news outlets released the Tsarnaev brothers' names, effectively resolving the confusion, and the Boston Bombing conversation soon veered away from Tripathi. On April 23, Sunil's body was discovered. He had died weeks before the bombings (Bidgood, 2013).

For Rumor #3, we coded tweets that contained either "Sunil" or "Tripathi." This dataset of 29,416 tweets had a much lower ratio of misinformation to correction—though still larger than Mendoza et al.'s (2010)—at only roughly 5:1 (22,819 misinformation to 4485 correction). In this case, the propagation of misinformation took a different shape, with misinformation compressed into shorter time window. Between 8pm on April 18 and 2am on April 19, tweets linking Tripathi to the bombing were sent at a rate of about 1-3 tweets per minute, most of which were speculative. Then, between 2:40am and 3am, misinformation

jumped from 40 tweets in ten minutes to 4690. Peak volume, however, was not sustained as it was for the other two rumors, declining at an (exponentially) steady rate, with occasionally mini-spikes.

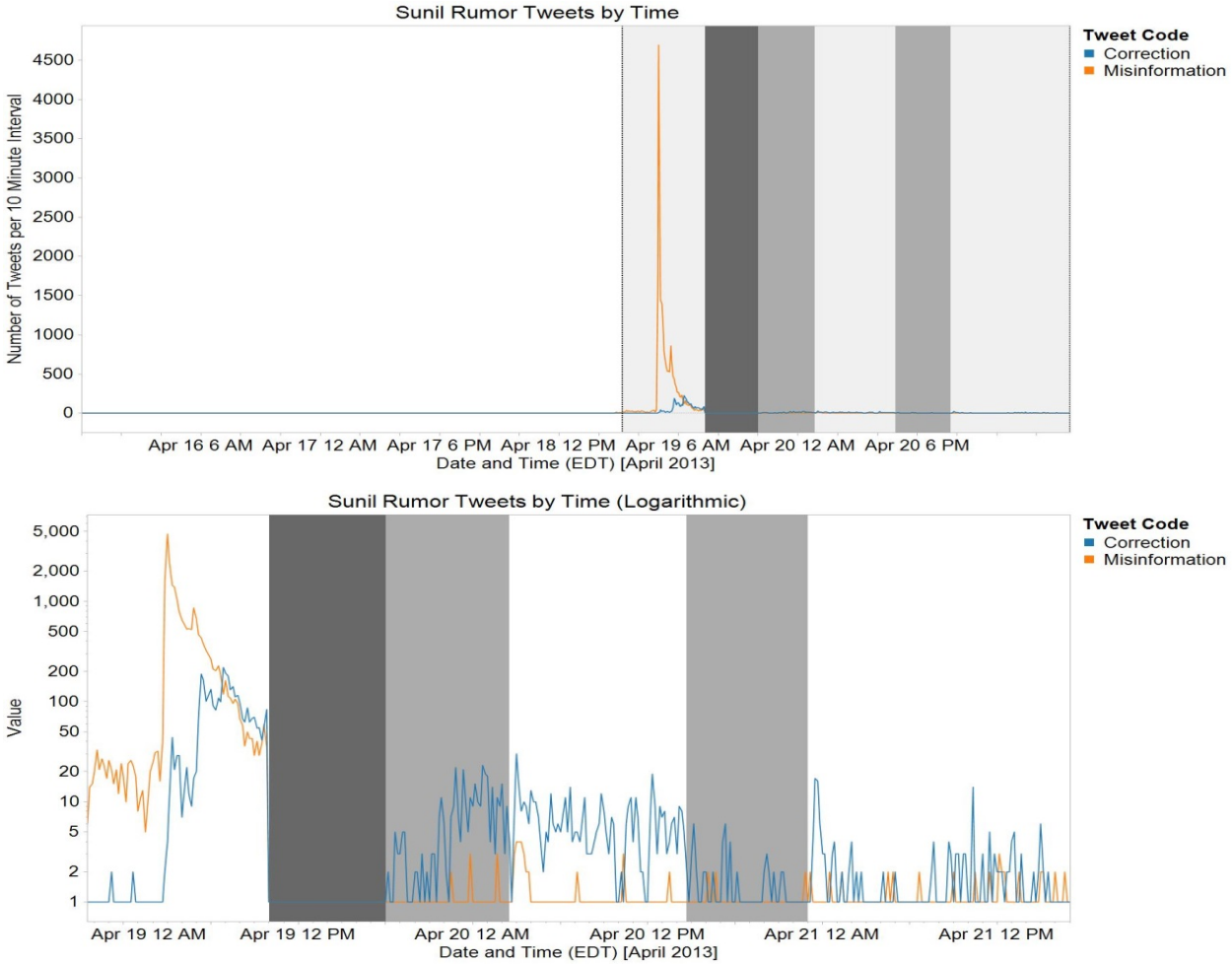

Figure 5. Sunil Tripathi Rumor, Tweets per 10 Minutes

*Light grey rectangle in top image highlights window of focus for bottom image

Remarkable here is the interaction between misinformation and correction. Again, corrections lagged behind misinformation in time and overall volume. However, they grew steadily and eventually overtook misinformation—soon after the official announcement naming the Tsarnaev brothers as suspects, which managed to effectively end the spread of misinformation within a few hours. Although data loss prevents a thorough analysis of how the rumor died off, we can see that corrections had already begun to overtake misinformation before the official announcement. For this rumor, corrections persist long after the misinformation fades as users commented on lessons learned about speculation.

5 Conclusion - Re-evaluating the Idea of the Self-Correcting Crowd

5.1 **Characterizing Misinformation**

Mendoza et al. (2010) demonstrated that the social media crowd has the potential to self-correct. Our study examines more closely the relationship between misinformation and corrections on Twitter. In support of Mendoza, we find evidence of crowd-correction for each rumor but with considerably smaller proportions of correction. Though misinformation and correction seem to rise and fall in tandem, they exhibit different magnitudes and a lag between the onset of misinformation and the correction. If we characterize these as patterns or signatures (Nahon, et al 2013), the frequency and wavelength of misinformation and correction are often aligned, but the amplitude can be exponentially different and there is often a delay in the correction signal. Additionally, in cases like Rumors 1 and 2, misinformation can persist at low levels that no longer activate significant corrections.

5.2 **Future Work**

This is preliminary work in a larger research effort on understanding the propagation of rumors through social media. We eventually would like to develop methods for automatically identifying misinformation by detecting the corrections. In the immediate future, we intend to analyze a larger set of rumors related to multiple crisis events. We hope to identify patterns or common types of rumors, possibly using "signatures" or characteristic patterns of misinformation and corrections over time. We intend to examine links within tweets—the URLs themselves and the domains to which they belong—to see if these features offer insight. Preliminary work suggests that tweets with misinformation contain *more* links than tweets with corrections, but that corrections tend to link to a higher number of different sources.

6 Acknowledgments

This work is a collaboration between the SoMe Lab the emCOMP Lab at UW and was partially supported by NSF Grants 1342252 and 1243170. The authors thank Shawn Walker for his extensive efforts to transform the original Twitter data into datasets that were more easily accessed and analyzed.

References

Alabama News, April 15th, 2013. Boston Marathon explosion: University of Mobile coach offers first-hand reports from scene. http://blog.al.com/live/2013/04/boston_marathon_explosion_univ.html

Bidgood, J. (2013). Body of Missing student at Brown is discovered. *NY Times*, (April 25, 2013) Available at: http://www.nytimes.com/2013/04/26/us/sunil-tripathi-student-at-brown-is-found-dead.html?_r=0

Castillo, C., Mendoza, M. & Poblete, B. (2011). Information Credibility on Twitter. In *Proceedings of on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 675-684.

Herrman, J. (2012). Twitter is a Truth Machine. Blog in *Buzzfeed*. (October, 2012). Available at: http://gofwd.tumblr.com/post/34623466723/twitter-is-a-truth-machine

Heverin, T. and Zach, L. (2012), Use of microblogging for collective sense-making during violent crises: A study of three campus shootings. *J. Am. Soc. Inf. Sci., 63*: 34–47.

Hill, K. (2012). Hurricane Sandy, @ComfortablySmug, and the Flood of Social Media Misinformation. *Forbes.com.* (October 30, 2012). Available at: http://www.forbes.com/sites/kashmirhill/2012/10/30/hurricane-sandy-and-the-flood-of-social-media-misinformation/

Hughes, A. L. and Palen, L. (2012). The Evolving Role of the Public Information Officer: An Examination of Social Media in Emergency Management, *Journal of Homeland Security and Emergency Management*, *9*(1), article 22.

Kotz, D. (2013). Injury Toll from Marathon Bombs Reduced to 264. *The Boston Globe*. (April 24, 2013). Available at: http://www.bostonglobe.com/lifestyle/health-wellness/2013/04/23/number-injured-marathon-bombing-revised-downward/NRpaz5mmvGquP7KMA6XsIK/story.html

Kwak, H., Lee, C., Park, H. & Moon, S. (2010). What is Twitter, a social network or a news media? *Intl. WWW Conference*, (Raleigh, NC, 2010), New York: ACM, 591-600.

Latonero, M. & Shklovski, I. (2011). Emergency Management, Twitter, and Social Media Evangelism. *International Journal of Information Systems for Crisis Response and Management, 3*(4): 1-16.

Madrigal, A. (2013a). Hey Reddit, Enough Boston Bombing Vigilantism. The Atlantic. (April 17). Available at: http://www.theatlantic.com/technology/archive/2013/04/hey-reddit-enough-boston-bombing-vigilantism/275062/

Madrigal, A. (2013b). It wasn't Sunil Tripathi: The Anatomy of a Misinformation Disaster. The Atlantic. (April 19). Available at: http://www.theatlantic.com/technology/archive/2013/04/it-wasnt-sunil-tripathi-the-anatomy-of-a-misinformation-disaster/275155/

Mendoza, M. Poblete, B. & Castillo, C. (2010). Twitter Under Crisis: Can We Trust What We RT? *In Proc. of 1st Workshop on Social Media Analytics (SOMA 2010)*. NY: ACM, 71- 79.

Nahon, K., Hemsley, J., Mason, R. M., Walker, S., & Eckert, J. (2013). Information flows in events of political unrest. *iConference 2013 Proceedings* (pp. 480-485). doi:10.9776/1325); Available: https://www.ideals.illinois.edu/bitstream/handle/2142/39165/259.pdf?sequence=4

Palen, L., & Liu, S. B. (2007). Citizen Communications in Crisis: Anticipating a Future of ICT-Supported Participation. *Proceedings of CHI 2007*. New York: ACM, 727-736.

Palen, L., Anderson, K. M., Mark, G., Martin, J., Sicker, D., Palmer, M., & Grunwald, D. (2010). A vision for technology-mediated support for public participation and assistance in mass emergencies and disasters. In Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference (Edinburgh, United Kingdom, April 14 – 16, 2010). Swinton, UK: *ACM-BCS Visions of Computer Science*, British Computer Society, 1-12.

Petrovic, S., Osborne, M., Mccreadie, R., Macdonald, C., and Ounis, I. (2013) *Can twitter replace newswire for breaking news?* In: *ICWSM 2013*, Boston, MA, USA.

Qu, Y., Huang, C., Zhang, P. & Zhang, J. (2011). Microblogging After a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake, *Proc of CSCW 2011*. NY: ACM, 25-34.

Reddit. (2013). Is missing student Sunil Tripathi Suspect #2? Thread on Reddit. (April 15/19, 2013). Available at: http://www.reddit.com/r/boston/comments/1cn9ga/is_missing_student_sunil_tripathi_marathon_bomber/

Starbird, K., & Palen, L. (2011). 'Voluntweeters': Self-organizing by Digital Volunteers in Times of Crisis, In *Proceedings of CHI 2011*. New York: ACM, 1071-1080.

The Federal Bureau of Investigation (FBI) (2013). News updates. FBI. Available at: http://www.fbi.gov/news/updates-on-investigation-into-multiple-explosions-in-boston

Watson, P. (2013). Potential Boston Bombing culprits and persons of interest identified? InfoWars. (April 17, 2013). Available at: http://www.infowars.com/boston-bombing-culprits-found

Table of Figures