How Information Snowballs: Exploring the Role of Exposure in Online Rumor Propagation

Ahmer Arif, Kelley Shanahan, Fang-Ju Chou, Yoanna Dosouto, Kate Starbird, Emma S. Spiro[^+]

HCDE, Information School[^], Department of Sociology[+]

University of Washington, Seattle WA, 98195

{ahmer, kshan, fjchou, ydosouto, kstarbi, espiro}@uw.edu

ABSTRACT

In this paper we highlight three distinct approaches to studying rumor dynamics—*volume*, *exposure*, and *content production*. Expanding upon prior work, which has focused on rumor volume, we argue that considering the size of the exposed population is a vital component of understanding rumoring. Additionally, by combining all three approaches we discover subtle features of rumoring behavior that would have been missed by applying each approach in isolation. Using a case study of rumoring on Twitter during a hostage crisis in Sydney, Australia, we apply a mixed-methods framework to explore rumoring and its consequences through these three lenses, focusing on the added dimension of exposure in particular. Our approach demonstrates the importance of considering both rumor content and the people engaging with rumor content to arrive at a more holistic understanding of communication dynamics. These results have implications for emergency responders and official use of social media during crisis management.

Author Keywords

Rumoring; information diffusion; information contagion; crisis informatics; Twitter; disaster response

ACM Classification Keywords

H.5.3 Groups & Organization Interfaces: collaborative computing, computer-supported cooperative work

INTRODUCTION

During crisis events, individuals increasingly utilize social media platforms to search for and disseminate event-related information, offer social support and resources to those affected, check on family and friends, and share eyewitness accounts [6,24,28]. Despite the widespread use of these platforms during crises, many have questioned the viability of utilizing information in such settings for increasing situational awareness due to the prevalence of misinformation (i.e. false or inaccurate information). In fact, emergency responders have singled out rampant misinformation as one of the factors contributing to their reluctance to use social media as a source of actionable information during crises [11,14,32]. Responding to these concerns, scholars continue to explore the dynamics of misinformation—and more generally, rumoring behavior—on social media, particularly during crisis events.

Classical social science research furnishes us with two important ways of understanding rumor prevalence: (1) in terms of the amount of rumor-related information present in the environment, and (2) in terms of the number of individuals who have encountered or heard a particular piece of information. However, prior studies of rumoring in online spaces have often taken a narrower view of rumor dynamics [1,3,7]. Specifically, studies of online rumoring during crisis events often privilege the first framing by focusing primarily on the magnitude of rumor-related content over time. For example, in their study of rumor transmission on the Chinese microblogging platform Sina Weibo, Liao and Shi [18] explore the total number of messages, before moving on to consider and distinguish message volume along content/user categories. Likewise, Starbird et al. [29] focus their analysis on message volume to identify critical moments in the rumor propagation during the Boston Bombing. Spiro et al. [28] also model the rate of posts over time in their exploration of rumoring during the Deepwater Horizon oil spill in 2011. Indeed, many recent studies on rumoring online focus primarily, or exclusively, on the raw magnitude of rumor-related content over time, i.e. rumor volume.

We argue that a volume perspective in isolation can miss important theoretical and practical dimensions of rumor dynamics such as the downstream resurgence of rumors. As such we reintroduce exposure as a vital component of rumoring, integrate it with the volume framing adopted by prior studies and apply this expanded perspective to an illustrative case study to investigate rumoring behavior online. In doing so, we are interested in identifying features of a rumor's temporal *signature* that offer insight into how individuals respond to and interact with rumor content. In particular, from the perspective of exposure, we explore how an individual's embeddedness in the larger social network may lead to specific rumor dynamics. We also consider how technical affordances of the system influence changes in the

CSCW '16, February 27-March 02, 2016, San Francisco, CA, USA © 2016 ACM. ISBN 978-1-4503-3592-8/16/02…$15.00 DOI: http://dx.doi.org/10.1145/2818048.2819964

information space, leading to increased or decreased prevalence of a particular rumor.

In the work presented here, we emphasize the importance of understanding both rumor content and the properties of the individuals who engage with such content. We argue that a more complete picture of rumoring behavior can be had by integrating these perspectives with the existing focus on total volume of messages. Our main contributions are the following:

- We expand on previously proposed approaches to quantifying rumor dynamics by emphasizing exposure to rumor content and illustrate the value of incorporating this perspective in an in depth case study.
- We characterize rumoring behavior on Twitter during a hostage crisis event using ideas of rumor *volume*, *exposure*, and *content production* to reveal how both message and user characteristics can change the information space surrounding rumors.
- We offer a typology for characterizing rumoring behavior along dimensions of volume and exposure to highlight phenomena critical to understanding rumor propagation.

RELATED WORK

Social Media and Crisis Events

Social media has become a critical component of crisis response and recovery, both from the perspective of official responders and the general public. As such, there has been increased interest from social and crowd computing scholars in understanding how people use microblogging sites and other social media platforms during extreme events: natural disasters, technological crises, confrontations, civil unrest, terrorist attacks, etc. [12,13,19,23,25,28,31,32]. Previous work on *crisis informatics* has shown that when crises occur, these platforms are appropriated by those affected (both directly and indirectly) to share information, offer social support and resources, and collectively make sense of the event [19]. Emergency responders are also increasingly turning to social media as a channel for communicating and interacting with their constituents since these tools offer new venues for people on the ground to share eyewitness accounts, which at times are at odds with official response plans and narratives [12].

Researchers and practitioners alike recognize the value of social media data as an information source during extreme, non-routine circumstances such as crisis events, yet many challenges remain. In particular, recent high profile events such as Hurricane Sandy and the Boston Marathon Bombing illustrate the potentially detrimental role that misinformation can play in this context [10]. Research on the spread of misinformation within and across social media during crisis events is still just beginning. However, sociological studies of the more general case of rumoring behavior have a longstanding tradition, and can offer important theoretical contributions to understanding how these phenomena unfold online.

Rumor Theory and Dynamics

In the social sciences, rumoring behavior is regarded as a social process of collective sensemaking through which individuals can understand situations characterized by high levels of uncertainty, anxiety and a lack of official news; it is precisely in these situations that rumors are likely to emerge [1,26,27]. Indeed, the crisis context was often used as a research environment in these early studies because it was ripe for rumors to spread [2,15]. Classical work on rumor dynamics points to characteristics that influence overall rumor prevalence, most notably the information's importance and ambiguity; that is, prominent rumors tend to pertain to salient, topical content but with a degree of uncertainty or anxiety [2]. Importantly, this definition of rumoring behavior does not rely on the veracity of content; instead, rumors are treated as information or stories of unknown validity.

A core aim in many explorations of rumor dynamics was to understand their spread or diffusion through a population. While these early studies discuss rumor prevalence in terms of information content, they more often conceptualize propagation through the number of individuals exposed to rumor content. They also primarily examine rumor transmission via face-to-face (FtF) interactions in naturalistic settings. This work suggested diffusion might be associated with both social and physical distance [7, 13,21,32]. More recently, many of these themes have been explored within social media spaces as well [20,28,31]. Collectively, these prior works emphasize theoretical frameworks for understanding rumor dynamics, and as such provide a valuable grounding for our research. However, much of this very early work suffers from a lack of empirical support. Though understandable due to the difficulty of studying rumors in informal FtT social networks, this motivates additional exploration of rumor dynamics.

METHODS

Background on Event Case Study

On December 15, 2014 a lone gunman took 18 people hostage inside the Lindt Chocolate Café at Martin Place in Sydney, Australia [35]. The resulting 16 hour standoff, which began at 9:45am AEDT, ended tragically with four individuals being injured and the deaths of two hostages and the gunman himself after police stormed the building bringing the crisis to a close. The siege was characterized by many uncertain circumstances around the gunman's motivations and police actions, resulting in a complex information space that spanned both traditional mass media outlets as well as social media platforms.

Data Collection

Our research team collected data during the Sydney Siege event for the explicit purpose of examining rumoring behavior during crisis events. Data collection utilized the Twitter Streaming API to track specific event-related terms

and phrases including: `sydneysiege`, `sydney`, `lindt`, `martinplace`, and `chocolate shop`. Data collection started on December 15 at 11:06am AEDT and ended two weeks later giving an observation window of 14 days over which a total of 5,429,345 tweets were archived for research purposes. This is the data we use in the remainder of the paper.

Identifying Rumors

To explore online rumoring behavior during crisis events, we first identify a set of rumors and their associated (i.e. relevant) tweets. For the purpose of this study, building on sociological formulations of rumoring, we treat a rumor as information or a story of unknown validity. Rumors can therefore turn out to be either true or false.

To identify rumors within the Sydney Siege corpus of tweets, we begin by consulting external sources (i.e. media reports) along with visual and text-based explorations of the Twitter data itself. Through an iterative process, rumor definitions are refined and then characterized by a specific search query designed to produce a low noise, comprehensive sample of tweets for each rumor. This strategy has been used with good effect in prior work on rumoring during crisis events [19]. In the case of the Sydney Siege data, our team identified five substantial rumors, three of which will be used as illustrative examples of the multi-perspective approach to rumor dynamics presented in this work[1].

Categorizing Tweets

To capture various aspects of rumoring activity, we categorize, or code, each distinct tweet in each of the rumor datasets via a manual, qualitative coding process. Our coding scheme includes the following five, mutually exclusive categories: Affirm, Deny, Neutral, Unrelated, and Uncodable. Coders evaluate the content of each tweet and assign one of these five codes. In the analysis that follows we focus on tweets coded as Affirm, Deny, or Neutral. A tweet coded as *affirm* means that it endorses or otherwise supports the given rumor whereas a tweet coded as *deny* disputes or refutes the rumor. A tweet coded as *neutral* neither directly affirms nor denies the rumor but is still related to the story.

Three trained coders manually code every distinct tweet in each rumor dataset (retweets and very close matches are removed for purposes of coding efficiency). For adjudication (i.e. when coders disagree), we apply a majority rules decision such that agreement by two or more coders determines the categorization. In the case where all three of the initial coders select distinct codes, we add a fourth coder to adjudicate. An inter-rater reliability analysis using the Fleiss' Kappa statistic [8] was performed to ascertain consistency among raters and found to be Kappa = 0.892 ($p < 0.001$).

Perspectives on Rumoring Behavior

To develop a richer conceptual understanding of how rumor dynamics occur in Twitter's information landscape during crisis events, we adopt (and advocate for) several different interpretive lenses. Each of these frames for exploring rumoring behavior is motivated by prior work within the social sciences [1,27], where studies focus on the number of people who have been exposed to a particular rumor. Unlike early work in FtF settings, however, studies of rumoring online stress the number of messages at any given time. Both areas of work have long been interested in serial transmission of rumors. We describe each of these perspectives, discussing their advantages and weaknesses. However, it is through combining these perspectives that we find additional insight that would have been missed by simply approaching our research question via one perspective.

The Volume Perspective

As many classical studies of rumoring suggest, the first interpretive lens through which one may explore rumoring behaviors is that of information or rumor *volume*. While prior work often uses notions of volume [16,19,20,22,28], studies rarely define their measure of volume. Volume can be quantified as the number of rumor-related social media posts observed in each discrete time interval. Here, we measure the volume, V, of the i^{th} rumor, r_i, at time t as the number of rumor-related tweets, m, observed at time t.[2]

$$V_{r_i t} = m_{r_i t}$$

Quantifying rumoring behavior through the lens of volume helps measure a rumor's overall prevalence in terms of the sheer number of messages present in the information space at a particular point. Rumor volume captures the magnitude of the online crowd's engagement in terms of observable actions—posting messages. Considering volume over time gives one view of how the rumor grows or shrinks in its lifetime. In particular, volume quantifies opportunities for individuals to encounter rumor-related tweets when searching the information space. As such, the volume perspective is well suited for questions about rumor magnitude over time, such as when does the rumor corpus reach its maximum size or how variable is the size of the rumor-related content over time.

Moreover, volume calculations lend themselves to additional refinements; for example, here we consider both volume of rumor affirming tweets and rumor denying tweets separately. This additional dimension of the volume perspective helps us to understand the general direction of conversation with respect to the public crowd's attempt at making sense of the

[1] We decided to focus on these three rumors specifically because they are less noisy and have clear boundaries. The other two rumors focused on speculation around the political affiliations of the gunman and the status of Sydney's airspace during the crisis.

[2] While this formulation assumes complete data about each rumor, volume (i.e. the rate of messages over time) can also be easily estimated if missing data exists.

rumor's content. Volume signatures serve as a useful starting point for our analysis of rumoring behavior, aiding in the selection of unique time-windows for more in depth analysis. Rumor volume also provides an overall frame of reference through which to orient rumoring phenomena, as will be seen in the results presented below.

The volume perspective also has limitations. Rumor volume is, by the definition above, agnostic to the number of individuals participating. In other words, observing a rumor volume of 10 messages could capture 1 tweet from each of 10 people or 10 tweets from 1 person—two very different cases. In addition, the researcher must choose a discrete time window over which to count (i.e. measure) volume. As will be seen subsequently, choice of an appropriate window may vary depending on the phenomena of interest; one must consider the timescale over which the phenomena of interest occur in order to choose a window size.

The Potential Exposure Perspective

Rumoring behavior is not simply a function of the number of messages related to the story, but also the number of individuals involved. Indeed much of the sociological studies discussed previously originally conceptualized rumor propagation in terms of the number of individuals who had heard a particular rumor. In social media platforms, however, this is a challenging aspect of rumoring to measure. Our second perspective on rumoring behavior aims to quantify the magnitude of the population *exposed* to rumor content rather than focusing on the magnitude of the content alone.

Measuring rumoring behavior in terms of the size of the exposed population is akin to measuring disease prevalence in the epidemiology field. As such, we define a rumor's exposure at time t, E_t, as the sum of the audience size a, for each rumor-related message posted at time t.

$$E_{r_i t} = \sum_{j=1}^{M_{r_i}} a_{jt}$$

Quantifying rumor exposure on Twitter presents some notable methodological challenges. Ideally, to measure the size of the exposed population, one would be able to measure the number of individuals exposed to each message. While this may be feasible on some social media platforms, it is not possible given the current limitations of Twitter's data API. Exposure on Twitter can potentially result from searching the public timeline of content, via following relationships or hashtag tracking, and through tweets embedded in external platforms. Unfortunately, many of these exposure mechanisms (e.g. tweets embedded in external platforms) are impossible to observe directly, and therefore extremely difficult to estimate. We can however estimate approximate exposure using an estimate of the audience size of the tweet author (i.e. follower count). Each tweet contains in its metadata the number of followers for the author's account at the time the tweet was posted, allowing this approximation of message exposure to be easily computed. In a subscription system like Twitter, we believe this to be a reasonable approach; however, we recognize its limitations. We have deliberately called this an approximate exposure because in reality it is neither an upper nor lower bound on actual exposure. It is not an upper bound because exposure could occur outside of following relationships, as noted. It is also not a lower bound because there is no guarantee that all followers actually view a particular message.

As with the volume perspective, exposure can be delineated by message attributes or categories. In our case, affirm versus deny distinctions are of particular interest. Estimating rumor exposure helps us to identify and compare posts from highly visible accounts versus those from less prominent accounts. Moreover, by considering the interaction between volume and potential exposure, we open spaces to interpret how different combinations of conditions might be impacting the information space in interesting ways. This might involve observing how a high potential exposure affirm tweet might be associated with a spike in volume occurring within the next minute but it could also point us to considering less observable effects. For instance, a high potential exposure denial tweet might help explain a sharp decrease in the volume of affirms if we consider that seeing a denial message might make some individuals less likely to tweet altogether rather than sending a message that takes a stance either way.

The Content Production Perspective

While both volume and exposure offer specific perspectives grounded in classical theories of informal communication on how to analyze rumor dynamics, each treats one tweet as equivalent to another tweet. However, in the context of social media platforms, retransmission or serial transmission of content is a prominent propagation mechanism and distinct from content produced by the original poster (i.e. author). While foundational work on rumoring also emphasized retransmission of content during crisis events [28,33], it was primarily because of the potential for distortion. Online, however, retransmission can be automated with the click of a button and allows for exact duplication of the original message. As such, distinguishing between original posts and re-posts is a particularly important component of rumoring behavior to measure.

The third approach leveraged here aims to build on the previous perspectives to estimate the impact of 'derivative content' generated by tweets [29]. We distinguish between *original content*—posts that are actually written/composed by the author, i.e. non-reposts, and *derivative content*—all identifiable instances of downstream content that are direct or very close copies of the original. This *content production* perspective utilizes metadata attached to each tweet by the Twitter API in combination with additional post-processing. Therefore, while data collected from Twitter makes this computation relatively easy, the phenomenon of serial transmission is not specific to the Twitter platform, and can be applied in many different social media domains.

Generally, tweets captured via the Twitter API are time-stamped and embedded with information about whether they are retweets (i.e. re-posts) of a particular tweet or not. This tag can be used to identify original content and establish an estimate of derivative content for those messages. In certain circumstances (e.g. when a tweet author used a third-party client, retweeted a link using an external website, or simply copy pasted a tweet's content to retweet 'manually'), this information can be inaccurate; thus, to address this limitation we strip all tweets of text tokens such as "@", "RT:" and URLs and chronologically order all copies, designating the first tweet as the original content and assigning all other copies and their derivative content values to this original tweet. These derivative content values are sometimes referred to as the *derivative volume* of a tweet.

We also calculate *derivative exposure*, which we define as the sum of the audience size (i.e. follower count) for each tweet included in an original tweet's derivative set. Although this value tends to be strongly correlated with the derivative volume in our data, it also gives voice to not just how many times the crowd engaged with an original message but also by whom, capturing important differences in the social embeddedness of these users.

Viewing rumor dynamics from the perspective of content production versus content derivation aids the researcher in understanding the realized impact of a particular tweet both in terms of how many times it was picked up by the crowd and how far it might have grown (or been distorted) compared to its initial footprint. It also teases apart two different rumoring behaviors: (1) introducing novel information into the environment, and (2) supporting and propagating information that already exists, both of which are important components of the overall dynamics of a particular story. As such, the content production lens is particularly suited to research that explores the similarities and differences between these two types of engagement.

Rumor Signatures

Expanding upon the work of Maddock et al. [19], each of the three perspectives described above offers a unique lens through which we can explore rumoring on social media during crisis events. Data related to the Sydney Siege will be used to illustrate the value of combining these multiple perspectives, leading to a more complete picture of rumor dynamics. For each measure of rumoring, we articulate or visualize the behavioral signatures over time, digging deeper into particular aspects of these signatures and supporting these results with additional data and qualitative insights.

FINDINGS

In this section we illustrate a multi-perspective approach to exploring rumor dynamics during the Sydney Siege hostage crisis. Our analysis takes the form of a case study, developing descriptions of three rumors in particular and discussing the implications of our approach for studies of rumoring on social media during crisis events. Unless specified otherwise, all times are in Australian Eastern Daylight Time (AEDT).

Rumor 1: The Lakemba Raids

During the Sydney Siege, one emerging rumor asserted that the Australian Federal Police carried out home raids in Lakemba, a predominantly Muslim suburb of Sydney, in parallel to, or as a response to, the hostage crisis. We refer to this rumor as the "Lakemba Rumor". This story turned out to be false, as authorities quickly denied these claims. Information suggests that the rumor may have originated from the sighting of twenty officers conducting a pre-arranged tour of the Lakemba Mosque as part of a police induction day.

We begin by considering rumor volume, separating out tweets in each of the three coding categories: affirm, neutral and deny, as well as distinguishing tweets via the content production classification, as seen in Figure 1.

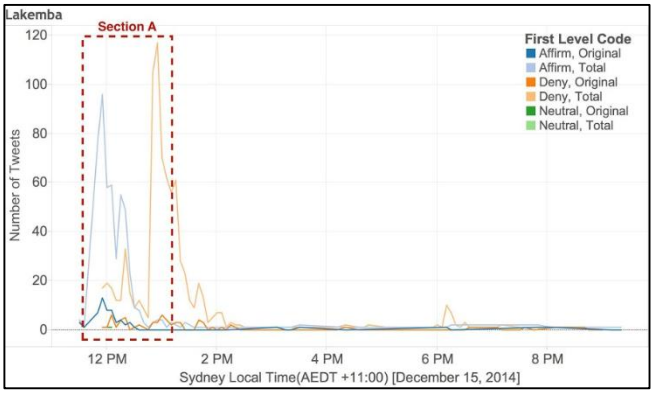

Figure 1. Lakemba Rumor Tweets by Time (10 minute intervals)

The highlighted and labeled Section A, Figure 1 (11:40am to 1:10pm) is a period of notable activity. Not only is volume at its highest in this window, but we also see a significant shift in response within the information space; the majority of tweets by volume change from affirming the rumor to denying during this window.

While volume is the starting point, it is through our multi-lens approach that we gain a better understanding of the rumor dynamics at play. In Figure 2, we compare volume and exposure for the period of time highlighted in Section A, Figure 1. Building from the basic rumor volume signature, shown in Figure 1, we now overlay the estimate

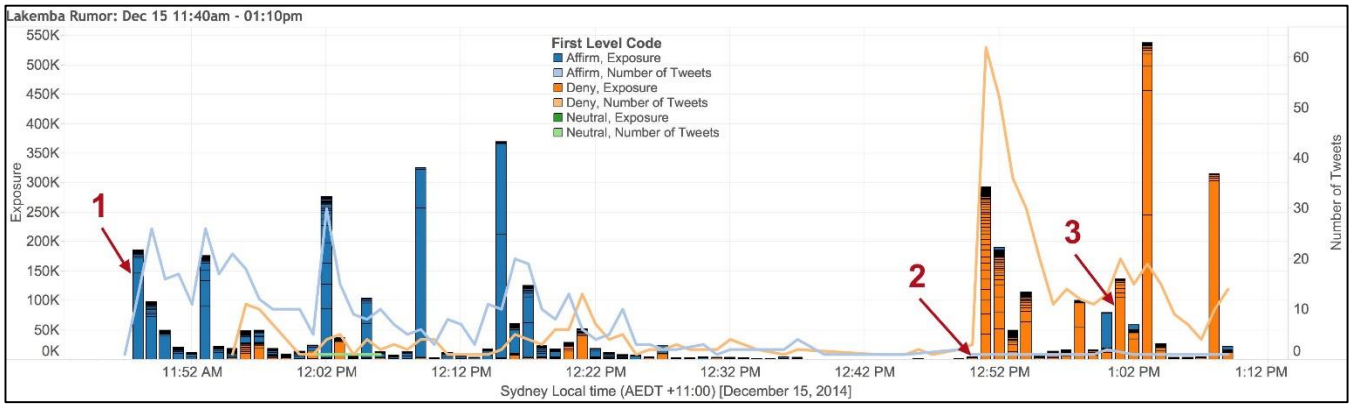

Figure 2. Lakemba Rumor Tweets and Potential Exposure by Time (1 minute intervals)

of rumor exposure. Each of the blue, orange, and green bars represents the total estimated exposure at a particular time. These bars are comprised of individual blocks (i.e. segments) each representing the audience size of one rumor-related tweet. Where the bars appear to be shaded black, the total exposed population is comprised of many tweet authors each with small audience size. In other cases, bars are visibly delineated into blocks, showing tweet authors with extremely large estimated audience size.

Immediately obvious in this representation of rumoring behaviors are points in time where the volume and exposures perspectives move in tandem, and points where they present very different interpretations of rumoring behavior. In particular, during this time frame we see three trends: (I) tweets with low exposure that are followed by high volume, (II) tweets with high exposure that are not followed by increased volume, and (III) tweets with high exposure that are followed by a surge in volume. We inspect each of these three cases in turn.

At 11:48am (Figure 2), a large news media outlet, Sky News Australia, entered the rumor by sending the following tweet:

@SkyNewsAust (11:48am): NSW Police + AFP are raiding several homes in Lakemba right now. More #martinplacesiege #sydneysiege #sydneycafesiege (@PMOnAir) [Point 1]

With 146,022 followers at the time, Sky News Australia seems to have had a significant impact in the information space. This particular tweet generated an estimated 137 instances of derivative content and had a derivative exposure of 370,280. Together these results suggest that high exposure at Point 1 provided the catalyst to the subsequent surge in the volume of affirmations related to the Lakemba rumor.

In Figure 2 we can also see a significant change in volume—shifting from majority Affirm to majority Deny tweets around 12:35pm. Inspection of the data reveals the following tweet, sent at 12:50pm by the Australian Federal Police (AFP).

@AFPMedia (12:50pm): Reports that the AFP is conducting search warrants in the Sydney suburb of Lakemba are incorrect. [Point 2]

After this post (Figure 2, Point 2), the volume of denial tweets began to surge, going from 3 tweets per minute (TPM) to 62 TPM at 12:51pm. At the time, the AFP's follower count was 2,569. The above tweet spurred an estimated 480 derivative messages and its derivative exposure is calculated to be 1,886,840.

This post illustrates a key example of an account with a relatively small audience size, and therefore low exposure, posting a tweet that leads to a large cascade—and high volume—of tweets. Perhaps more importantly, the crowd couples this behavior with a shift in overall opinion about the rumor (denials take over affirmations).

Lastly, consider Figure 2, Point 3:

@TweeterA (1:01pm): Rumors abound. #SydneySiege RT @AFPmedia Reports that the AFP is conducting search warrants in the Sydney suburb of Lakemba are incorrect. [Point 3]

This individual's tweet is a clear example of information with high exposure but low derivative volume. The follower count at the time was 105,106, but the tweet only produced 6 recorded instances of downstream derivative content and had a derivative exposure of 20,014. While this post had a high initial exposure, it did not make it far in the information space of this rumor due to lack of serial transmission, suggesting it had a very low impact on the rumor's overall propagation.

Rumor 2: Ray Hadley Speaks to Hostages

Our second case of interest centers on claims that an Australian radio host named Ray Hadley had spoken off-air to a hostage during the Sydney Siege standoff. This story was later confirmed to be true. We shall refer to this story as the "Hadley Rumor" in our discussion.

As with the Lakemba Rumor, we approach this second illustrative case via the three perspectives outlined above. Figure 3 illustrates rumor volume, parsing out original

content and separating the overall rumor signature into affirmations, neutral tweets, and denials. As with the first case, we highlight a specific time frame during which the majority of rumoring behavior took place and look at this time frame in more detail through the multi-perspective approach.

The Hadley rumor had one 'big moment' when volume peaked at 79 TPM, as highlighted in Section A in Figure 3 between 1:15pm and 2:20pm. As seen in Figure 4, viewed through the multi-lens approach, we find that tweets during this period have distinct initial exposure classifications: low exposure tweets and high exposure tweets.

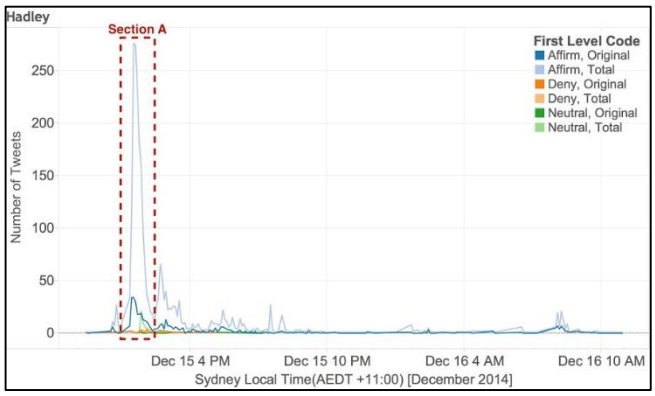

Figure 3. Hadley Rumor Tweets by Time (10 minute intervals)

Just before the highest peak, at 1:29pm, the Australian talkback radio station where Ray Hadley worked, 2GB 873 AM, sent the following tweet (Figure 4, Point 4).

@2GB873 (1:20pm): Ray Hadley is taking an extended ad break as he's talking on the phone to a possible hostage. Listen live http://t.co/2MMBPUgplE [Point 4]

At the time, 2GB873 had 14,947 Twitter followers. We consider this tweet a medium exposure tweet. However, despite not having high exposure, it generated a large cascade in rumor-related conversation volume. Nine minutes later, rumor volume peaked with 79 TPM at 1:29pm. Relatedly, this tweet from 2GB873 had a derivative content value of 57, which suggest that its content caught the attention of the public audience on Twitter during this time. Here, we illustrate a case where low exposure produces an increase in volume through multiple individual actions by downstream accounts passing along and amplifying the original message.

Throughout this time, there were also multiple accounts with high follower counts involved in the conversation. Several intermittent smaller peaks followed the largest peak in estimated exposure; by 1:47 pm, the conversation volume was consistently at or below 31 TPM. At this point we observe The Australian, the biggest-selling national newspaper in the country [36], posting the following tweet (Figure 4).

@australian (1:48pm): 2GB's Ray Hadley has been contacted by a #SydneySiege hostage. http://t.co/XDiHhoX95k [Point 5]

At the time, The Australian had 228,209 Twitter followers, one of the highest initial exposures for a single tweet in our dataset. Within minutes, several accounts picked up the @australian tweet and re-posted it. This tweet had a derivative volume of 28, suggesting that the @australian post actually led to a secondary peak in exposure that corresponds to the same interval in which the original tweet was shared. The audience size of that account could be entirely responsible for the downstream spike, illustrating the power of a single tweet by a highly visible account.

This particular example constitutes a high exposure tweet that was able to contribute to the increase or maintenance of rumor volume in the information space. Both sample tweets shown in this section contributed to an increase in overall volume and were able to generate high derivative content.

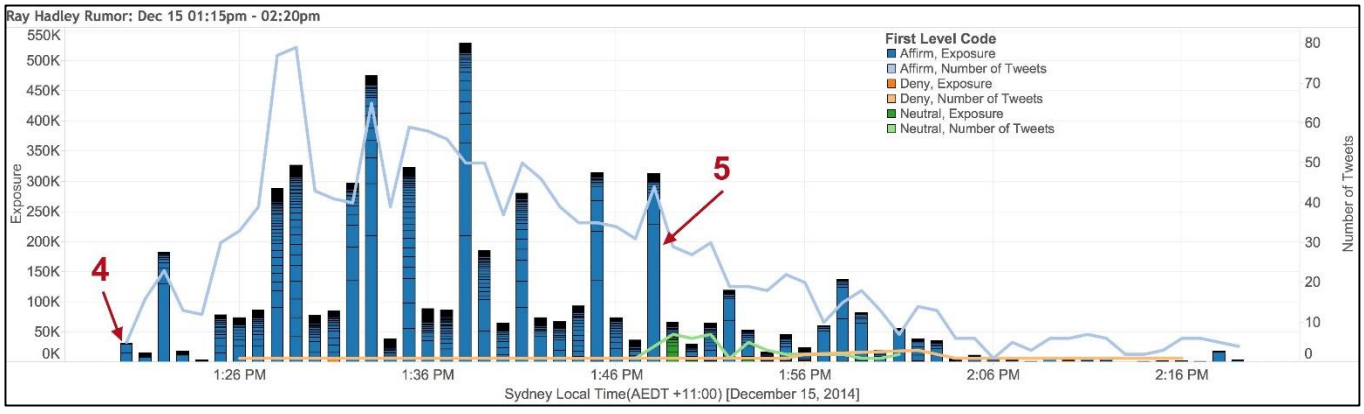

Figure 4. Hadley Rumor Tweets and Potential Exposure by Time (1 minute intervals)

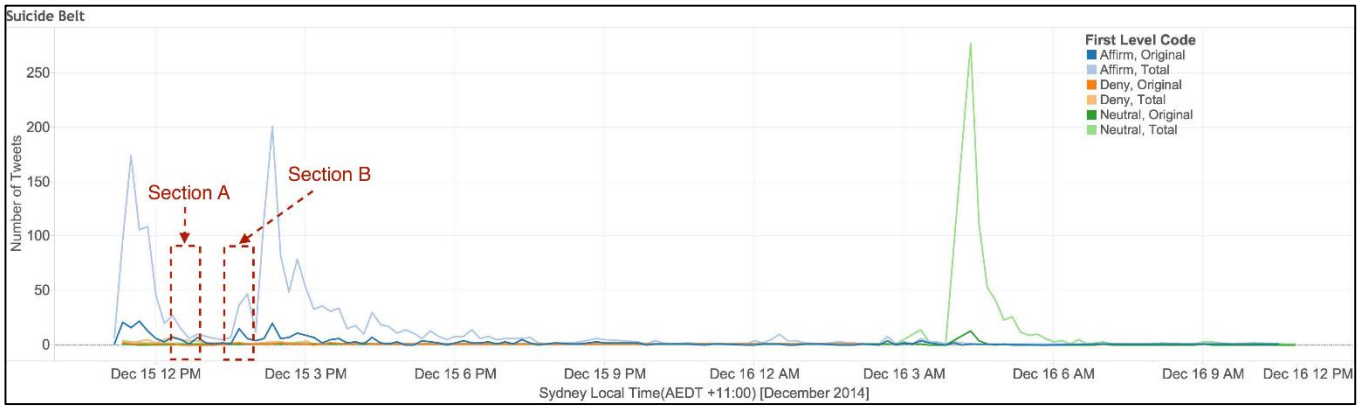

Figure 5. Suicide Belt Rumor Tweets by Time (10 minute intervals)

Rumor 3: Suicide Belts

During the Sydney Siege, there were claims that hostages or hostage takers were spotted wearing backpacks or some sort of wearable accessory that could contain an explosive device. This sighting generated a rumor that asserted the presence of suicide belts or vests during the siege. This was later confirmed to be false information. We refer to this at the "Suicide Belts Rumor."

As is familiar, we begin our exploration by considering rumor volume over time, distinguishing between notable categories and content production types. In Figure 5, we see that the rumor has two separate spikes in affirming tweets followed much later by an even larger spike of neutral tweets. This final spike consisted of largely informational tweets conveying details learned after the siege ended:

@TweeterB (6:21pm): #SydneySeige MT @cnnbrk: Source: #Sydney hostage-taker was killed; he wore thick black vest & authorities checked for explosives.

The areas highlighted in Figure 5 (Sections A and B) again showcase moments with unique behavior due to the presence of multiple ripples. Even though these sections do not offer a high volume when compared to other areas of the rumor, it does offer an opportunity to explore multiple ripples in what may look like a more silent/non-active time window. Figure 6 drills into the tweets captured between 11:10am and 1:59pm where we find a collection of small echo waves produced after the first large wave of tweets.

@NewsOnTheMin (11:24am): MORE: One of the terrorist inside the coffee shop is wearing backpack and vest, likely a bomb. #Sydney http://t.co/88FHhLw3qo [Point 6]

This tweet (Point 6) is classified as an affirm tweet with a low initial exposure (at the time @NewsOnTheMin had 4,101 followers). Despite its low exposure the tweet inspired 39 instances of derivative content. Accounts with both large and small follower counts picked up the tweet and propelled it forward. Eventually the tweet reached an estimated derivative exposure of 279,175, which might have contributed to an increase in rumor volume. Not long after, ITV News, a British television program, tweeted the following:

@itvnews (11:27am): Sydney terror suspect 'is wearing backpack which could contain explosives' http://t.co/zQH8o3CxRI [Point 7]

This tweet (Point7) was classified as affirming the rumor; it had an extremely high initial exposure since 522,238 users followed the account at the time of collection. The tweet was quickly picked up by multiple individuals with high and low follower counts alike and had a derivative volume of 90. This downstream content contributed to an increase in volume that is depicted as the highest spike in Section A of Figure 5.

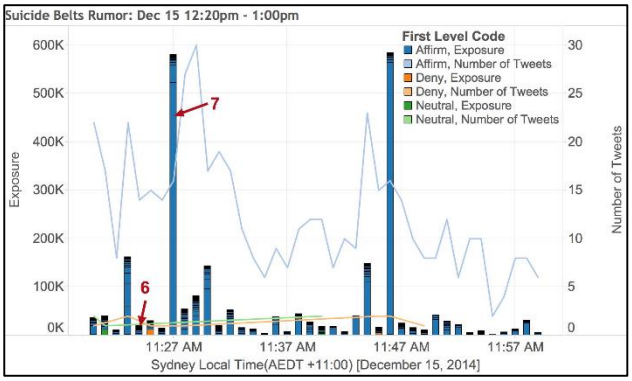

Figure 6. Suicide Belts Rumor Tweets and Potential Exposure by Time – Section A (1 minute intervals)

This tweet is an example of how high exposure can contribute to an increase in overall rumor volume by inspiring derivative content. Interestingly, even though Point 7 highlights a case with a disproportionately higher initial exposure than Point 6, both posts appeared to be salient to the Twitter audience, contributing to an increase in derivative volume and the total volume of the event.

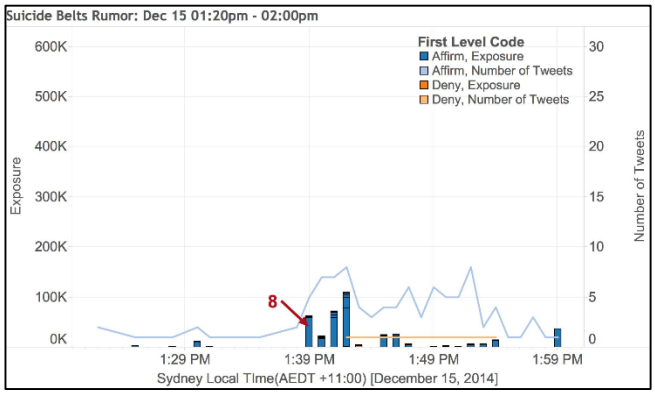

Figure 7. Suicide Belts Rumor Tweets and Potential Exposure by Time – Section B (1 minute intervals)

Two hours after Point 7, @SputnikInt, a news agency and radio station, produced the following (Figure 7):

@SputnikInt (1:39pm): Hostage-takers carrying 'suicide belts' take hostages, demand to meet with Australian PM in Sydney Siege: reports http://t.co/64rzbxodKp [Point 8]

The above tweet constitutes another affirm in this rumor's collection. Interestingly, however, it has the opposite effect from the other two tweets highlighted in this section. Unlike its predecessors (posted by @NewsOnTheMin and @itvnews), this tweet from @SputnikInt fails to capture the attention of its public, even when the author had an initial high exposure with over 59,000 followers. This tweet had a derivative content value of only two tweets and the overall rumor volume level after publication did not increase substantially, as indicated in Figure 7.

DISCUSSION

In the results presented above, we noted several patterns visible when volume was coupled with initial potential exposure and derivative content. If we enumerate the four idealized categories that result from a simple binary distinction between high versus low derivative volume, and likewise high versus low exposure, we find that each observed phenomena can be classified within this framework. Moreover, in synthesizing our results, this theoretical construct of derivative volume-exposure paired effects can be used to frame a discussion of the implications and insight of bringing exposure back into the study of rumor dynamics. Here, we utilize this construct—seen in Table 1—to draw connections among these patterns.

	High Derivative Volume	**Low Derivative Volume**
High P. Exposure	Giant effect	Fizzle effect
Low P. Exposure	Snowball effect	Babble effect

Table 1. Theoretical Exposure-Derivative Volume Paired Effects

To operationalize what is considered high versus low in our context, we calculate the mean derivative volume and exposure for all original content associated with each rumor. While this is a simple (and relatively coarse) measure, it proves to be effective in systematically classifying rumor effects such that important consequences can be discovered. We note, however, that appropriate distinctions between high and low volume/exposure may be context specific, and researchers should choose a threshold appropriate to the case of interest and data used. Applying this technique to the Sydney Siege data expands our understanding of rumoring behavior. We discuss each of the four named effects in Table 1 in turn.

Giant Effect

During segments of high exposure and high volume, we observed that rumoring is being driven by what we call the "giant effect". In these cases, rumor-related information is not only abundant but mostly derived from the posts of individuals with very high direct reach. Entering the information space with large initial footprints, these messages spur the creation of derivative content that shifts the overall volume signature of a rumor in noticeably significant ways. We often see official media and news outlets as big players in the giant effect category. An example of this comes from Figure 6, Point 7 in the Suicide Belts rumor when ITV News tweeted and spurred a surge in volume.

Rumor content during such periods seems to be salient to the public and therefore reposted. Interestingly, giant effects are not restricted to rumor affirmations or denials specifically, but can result in both cases. Furthermore, although giant effects may be triggered from prominent actors such as mass media accounts or celebrity figures posting salient content, they can also emerge from the emergency responder community. Indeed, previous work in this area has demonstrated that local government officials often become targets for public attention during crisis events [34].

Snowball Effect

We've categorized messages with low initial exposure and high derivative volume that drive increases in overall rumoring as "snowballs". Like a small ball of snow rolling down a snow-covered hillside that gains momentum, mass and surface area as it rolls along, these messages begin from an initial state of small significance and build upon themselves to become an information avalanche. Snowballs are particularly interesting in this context because they reveal instances where the volume and exposure perspectives offer differing measures of rumoring.

Often in crisis events, tweets from accounts with low follower counts get picked up by the crowd and diffuse at a large scale. For example, at Figure 2, Point 2 in the Lakemba rumor the Australian Federal Police's account tweeted a correction of the false rumor. After this point, affirmations significantly declined and denials rose. This tweet's initial

exposure was fairly low but it generated a high volume of derivative content.

Another general trend the snowball effect supports is the idea that during crisis events Twitter users arrive at information through sources other than their social network ties, such as through search functions or external articles. This corroborates The Million Follower Fallacy presented by Cha et al. [5]. In the context of crisis response and recovery, this phenomenon has important implications. While emergency responders may initially have low exposure, they may be able to design content aiming for snowballs. Moreover, exposure can be very dynamic and emergency responders could easily turn into giants overnight when extreme events occur [34].

Fizzle Effect

The third effect has been named the "fizzle effect" and captures moments where there was high exposure and low volume. These incidents occur when accounts with high followers tweet information that is not repeated or spread by others, leading to low derivative volume. These messages fizzle out and are lost within the larger stream. As with snowball effects, fizzle effects highlight the differences between rumor volume and rumor exposure.

It is also interesting to note when this effect tends to occur during the course of a rumor. We often see fizzles at the end of the signature, when the information space is dying out and activity has declined. Additionally, we see it during the crossover of Affirmations and Denials within rumors. This suggests that while a fizzle tweet may not increase volume, it may prevent an individual from tweeting false information. More specifically, a user may be exposed to a given tweet and instead of tweeting the same information, he/she may just not tweet the opposing information. We see this in the Lakemba rumor when the following tweet was sent out:

@TweeterC (12:01pm): Be wary of reports about police raids in Lakemba, Australia -- not true: http://t.co/bFLkjEU9Fh

Right around this time is when denials were introduced into the information space. Though the tweet fizzled, it is possible that individuals saw the tweet and while they did not pass along the information, they may have stopped from tweeting an affirmation.

Babble Effect

Our last derivative volume-exposure paired effect occurs during cases of low exposure and low volume. An account with low follower count tweets and the subsequent rumor volume remains low. This activity conjures images of shouting into the dark, or in this case, the general hum or incessant din of social media streams. Information is being added to the larger conversation but the source of that information has limited reach to others and content is insufficient to spark changes in volume. In the context of emergency response organizations, babble effects could be detrimental to response activities, because crisis-related information is much less likely to reach members of the public through a diffusion mechanism.

Summary

Rumor dynamics are a variable and complex example of collective behavior. On social media, especially during crisis events, rumors may differ by context and case, raising numerous challenges in measuring and modeling such behavior. As such, research frameworks that contribute generalizable theoretical constructs are rare. Not only are the expanded rumor signatures presented here advantageous in their simplicity, but they also allow the researcher to pull out concrete ideas matching qualitative features of rumor dynamics (e.g. our four volume/exposure effects) from very complex data, bringing important trends to light. Though here we demonstrate our approach using data from Twitter, these constructs can apply across many social media platforms. Indeed any environment that brings together people, information messages, and re-shares (e.g. Facebook, Pinterest, Instagram) can fit within our derivative volume-exposure pairs.

In the discussion above, we emphasize the implications of our results for emergency responders through each rumor effect. These implications are two-fold. First, being aware of the different types of effects a post may produce can help to design and direct crisis-related information. For example, making use of high potential exposure accounts combined with information designed for serial transmission could result in giant effects. Second, our approach highlights the fact that emergency responders should consider both messages and people when engaging with rumors. For example, misinformation posted by high exposure accounts could be more detrimental than posts by low exposure accounts. Moreover, it is important for emergency responders to understand how potential exposure may change over the course of the event.

FUTURE WORK

There are many ways to enhance the scope and applicability of the multi-lens approach to rumoring that we present here. Firstly, we intend to apply these constructs and replicate this analysis across a larger set of rumors and multiple events to evaluate this approach against more contexts. We also recognize that the measures used here are (purposively) coarse estimates designed to serve as simple starting points for qualitative interpretations and we intend to refine these significantly in the future. For instance, rather than computing values across the lifespan of rumors we intend to focus on more localized time windows around tweets when evaluating their impact. Together, these steps will help us evolve our understanding of the links between rumor growth and the impact of particular tweets and/or individual accounts. We hope that this in turn will enable us to build deeper and more robust constructs that inform our greater research agenda of rapidly detecting rumors on social media platforms in the future.

CONCLUSION

In this research, we adopt a mixed-methods approach for examining how individual messages related to rumors propagate on social media during crisis events. We identify three complementary perspectives – *volume*, *exposure* and *content production* – that integrate qualitative and visual interpretation with quantitative measures to highlight how these messages might affect overall rumor dynamics. We pay particular attention to incorporating estimates of rumor exposure as an important dimension of rumor dynamics that has been overlooked in prior work. To better ground and evaluate these approaches, we conduct an empirical case study where we analyze three rumors that spread on Twitter during the 2014 Sydney Siege. Our findings led us to articulating a conceptual construct that can inform future analyses in this area by exposing trends and helping generate new types of signatures for how rumors propagate.

ACKNOWLEDGMENTS

This research is a collaboration between the emCOMP lab and DataLab at the University of Washington and was supported by NSF Grants 1342252 and 1420255. We also wish to thank the UW SoMe Lab for providing infrastructure support as well as the students who provided significant assistance to this project, including John Robinson and Jim Maddock (among others).

REFERENCES

1. Gordon W. Allport and Leo Postman. 1947. The Psychology of Rumor. Oxford, England: Henry Holt.
2. Susan Anthony. 1973. Anxiety and Rumor. *The Journal of Social Psychology* 89, No. 1, 91-98.
3. Allen H. Barton. 1969. Communities in Disaster: A Sociological Analysis of Collective Stress Situations. Garden City, New York: Doubleday and Company, Inc.
4. Theodore Caplow. 1947. Rumors in War. *Social Forces*, 298-302.
5. Meeyoung Cha, Hamed Haddai, Fabricio Benevenuto, and Krishna P. Gummadi. 2010. Measuring User Influence in Twitter : The Million Follower Fallacy. *International AAAI Conference on Weblogs and Social Media*, 10–17. http://doi.org/10.1.1.167.192
6. William J. Corvey, Sarah Vieweg, Travis Rood, and Martha Palmer. 2010. Twitter in Mass Emergency: What NLP Techniques can Contribute. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media. Association for Computational Linguistics*, June, 23–24. http://doi.org/10.1186/1750-1326-4-10
7. Stuart C. Dodd. 1953. Testing Message Diffusion in Controlled Experiments: Charting the Distance and Time Factors in the Interactance Hypothesis. *American Sociological Review,* 18(4): 410-416.
8. Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*, 378–382. http://doi.org/10.1037/h0031619
9. Bradley Greenberg. 1964. Diffusion of News of the Kennedy Assassination. *Public Opinion Quarterly* 28, 2, 225–232. http://doi.org/10.1086/267239
10. Kashmir Hill. 2012. Hurricane Sandy, @ ComfortablySmug, and the Flood of Social Media Misinformation. *Forbes. Com*.
11. Starr R. Hiltz, Jane Kushma, and Linda Plotnick. 2014. Use of Social Media by US Public Sector Emergency Managers: Barriers and Wish Lists. *In Proceedings of ISCRAM* (2014).
12. Amanda L. Hughes, Lise A. A. St. Denis, Leysia Palen, and Kenneth M. Anderson. 2014. Online public communications by police & fire services during the 2012 Hurricane Sandy. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 1505-1514.
13. Amanda L. Hughes and Leysia Palen. 2009. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management 6*, 248. http://doi.org/10.1504/IJEM.2009.031564
14. Amanda L. Hughes and Leysia Palen. 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.
15. Jean-Noël Kapferer. 1989. A Mass Poisoning Rumor in Europe. *Public Opinion Quarterly* 53, 4, 467–481. Retrieved from http://www.jstor.org/stable/2749354
16. Kwon Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. *In IEEE 13th International Conference on Data Mining, (ICDM), 1103-1108.*
17. Kristina Lerman, Rumi Ghosh, and Tawan Surachawala. 2010. Social Contagion : An Empirical Study of Information Spread on Digg and Twitter Follower Graphs. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 90–97.
18. Qinying Liao and Lei Shi. 2013. She gets a sports car from our donation. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, 587. http://doi.org/10.1145/2441776.2441842
19. Jim Maddock, Kate Starbird, Haneen Al-Hassani, Daniel E. Sandoval, Mania Orand, and Robert M. Mason. 2015. Characterizing Online Rumoring Behavior Using Multi-Dimensional Signatures. *In*

Proceedings of 18th ACM Computer Supported Cooperative Work and Social Computing (CSCW'15).

20. Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter Under Crisis: Can we trust what we RT? *Workshop on Social Media Analytics*, 9. http://doi.org/10.1145/1964858.1964869
21. Andrew Noymer. 2001. The transmission and persistence of "urban legends": Sociological application of age- structured epidemic models. *The Journal of Mathematical Sociology 25*, 299–323. http://doi.org/10.1080/0022250X.2001.9990256
22. Onook Oh, Kyounghee H. Kwon, and H. Raghav Rao. 2010. An exploration of social media in extreme events: rumor theory and twitter during the Haiti earthquake. *International Conference on Information Systems (ICIS)*, 231.
23. Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 2015. What to Expect When the Unexpected Happens: Social Media Communications Across Crises. In *Proceedings of 18th ACM Computer Supported Cooperative Work and Social Computing (CSCW'15),* No. EPFL-CONF-203562. 2015.
24. Leysia Palen, Kenneth M. Anderson, Gloria Mark, James Martin, Douglas Sicker, Martha Palmer, and Dirk Grunwald. 2010. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In *Proceedings of the 2010 ACM BCS Visions of Computer Science Conference*, 1–12.
25. Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying Misinformation in Microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1589–1599.
26. Joseph Scanlon. 2007. Sampling unknown universe: Problems of researching mass casualty incidents (a history of ECRU's field research). *Statistics in Medicine 26*, 1812–1823. http://doi.org/10.1002/sim.2800
27. Tamotsu Shibutani. 1969. Improvised News: A Sociological Study of Rumor. *American Sociological Review 34*, 781. http://doi.org/10.2307/2092353
28. Emma S. Spiro, Jeannette Sutton, Matt Greczek, Sean Fitzhugh, Nicole Pierski, and Carter T. Butts. 2012. Rumoring During Extreme Events: A Case Study of Deepwater Horizon 2010. In *Proceedings of the ACM Web Science 2012 Conference (WebSci12)*, 275–283. http://doi.org/10.1145/2380718.2380754
29. Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M. Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 Boston marathon bombing. In *Proceedings of iConference 2014*, 654-662.
30. Kate Starbird, Leysia Palen, Amanda L. Hughes, and Sarah Vieweg. 2010. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 241-250. http://doi.org/10.1145/1718918.1718965
31. Kate Starbird and Leysia Palen. 2012. (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. In *Proceedings of the ACM 2012 conference on computer supported cooperative work,* 7–16. http://doi.org/10.1145/2145204.2145212
32. Jeannette Sutton, Leysia Palen, and Irina Shklovski. 2008. Backchannels on the Front Lines: Emergent Uses of Social Media in the 2007 Southern California Wildfires. In *Proceedings of the 5th International ISCRAM Conference*, 1–9.
33. Jeannette Sutton, Emma S. Spiro, Britta Johnson, Sean Fitzhugh, Ben Gibson, and Carter T. Butts. 2013. Warning Tweets: Serial Transmission of Messages During the Warning Phase of a Disaster Event. *Information, Communication & Society* 17, 6, 765–787. http://doi.org/10.1080/1369118X.2013.862561
34. Jeannette Sutton, Emma S. Spiro, Britta Johnson, Sean Fitzhugh, Ben Gibson, and Carter T. Butts. 2014. Terse message amplification in the Boston Bombing Response. In *Proceedings of the 11th International ISCRAM Conference*, 612-621.
35. Sydney Siege: How the Hostage Drama Played Out. The Sydney Morning Herald. Accessed May 13, 2015. http://www.smh.com.au/nsw/sydney-siege-how-the-hostage-drama-played-out-20141220-12b1km.html
36. The Australian. (n.d.). Retrieved May 4, 2015, from http://en.wikipedia.org/wiki/The_Australian