A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors

Ahmer Arif, John J. Robinson, Stephanie A. Stanek, Elodie Fichet[+], Paul Townsend, Zena Worku, Kate Starbird

HCDE, Department of Communication[+]
University of Washington, Seattle WA, 98195
{ahmer, soco, fjchou, stanek14, efichet, pjt33, zenaget, kstarbi}@uw.edu

ABSTRACT

This paper examines how users of social media correct online rumors during crisis events. Focusing on Twitter, we identify different patterns of information correcting behaviors and describe the actions, motivations, rationalizations and experiences of people who exhibited them. To do this, we analyze digital traces across two separate crisis events and interviews of fifteen individuals who generated some of those traces. Salient themes ensuing from this work help us describe: 1) different mechanisms of corrective action with respect to who gets corrected and how; 2) how responsibility is positioned for verifying and correcting information; and 3) how users' imagined audience influences their corrective strategy. We synthesize these three components into a preliminary model and explore the role of imagined audiences—both who those audiences are and how they react to and interact with shared information—in shaping users' decisions about whether and how to correct rumors.

Author Keywords

Social media; social computing; Twitter; rumoring; crisis informatics; imagined audience; folk theories

ACM Classification Keywords

H.5.3 [Information Interfaces & Presentation]: Groups & Organization Interfaces - Collaborative computing, Computer-supported cooperative work; K.4.2 Social Issues

INTRODUCTION

Finding and disseminating information rapidly can be crucial to building situational awareness during periods of collective stress and uncertainty. This is particularly true for crisis events, where activities such as offering support, learning from eye-witness accounts, and checking in with loved ones can help people both physically and mentally [12,30]. Consequently, social platforms such as Twitter that allow people to quickly communicate with a wide audience have come to be seen as an important medium for collective sensemaking activities during crises.

However, the conditions of disasters can also give rise to another closely related human response—the propagation of rumors. In these cases, the affordances of social media platforms serve to rapidly spread unverified information or even misinformation. This can have negative consequences for the efforts of emergency responders and the general public. Reflecting these concerns, both mainstream media and emergency officials have called out the threat of misinformation as limiting the utility of social media as a source of actionable information [19,26].

A growing body of research on the dynamics of online rumoring and detecting rumors on social media attests to the continued interest of disaster scholars in addressing this challenge. One aspect of online rumors that has drawn attention pertains to how these rumors are ultimately "self-corrected" by the online crowd. For example, Mendoza et al. find evidence suggesting that the online crowd posts more questions or challenges when presented with a false rumor [28]. Zhao et al. build upon this idea by examining how rumors might be identified by seeking out and clustering "enquiry" tweets that express skepticism about a story [44]. Examining this further, Starbird et al. measured the volume of corrections across the lifespan of several rumors and found them to constitute a relatively small portion of a rumor's overall propagation [38].

As such, digital traces of the crowd that question, deny or otherwise 'correct' rumors have been a focus of study for some time now. However, most of this research concentrates on leveraging these traces to detect online rumors. Consequently, very few studies have moved past the analysis of digital traces to interact directly with the users who were responsible for creating them.

We argue that this has led to an important gap in the literature when it comes to developing a deeper understanding of *how* members of the crowd choose to correct online rumors, and the different strategies they use to do this work. To address this, we examine two false rumors from two different crisis events through combined analysis of trace data (tweets) and interviews with 15

CSCW '17, February 25-March 01, 2017, Portland, OR, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4335-0/17/03…$15.00
DOI: http://dx.doi.org/10.1145/2998181.2998294

individuals who left some of these traces—specifically those who engaged in rumoring and correcting behaviors during these events. In doing so, we are interested in developing a deeper understanding of the drivers of and barriers to different rumor-correction behaviors.

BACKGROUND AND RELATED WORK

Social Media use during Crisis Events

A growing body of research attests to the widespread adoption of social media and the increasingly critical role they are playing in facilitating information-sharing during crisis events—from natural disasters like earthquakes [1,16,39] to man-made crises such as acts of terrorism [10]. Summarizing the breadth of this literature, in the wake of disaster events, people are repeatedly turning to these platforms with several purposes: to share information about their own circumstances, to seek actionable information for their own response actions and about impacts to the people and places they care about, to request and offer assistance, and to seek and provide emotional support to others. Increasingly, to be part of this conversation, emergency responders are integrating social media into their formal work practices as well. Though the capacity of social media to facilitate information-sharing practices has been described, in general, with a great deal of optimism, these platforms also bring new challenges. One commonly noted weakness of social media use in the crisis context is the vulnerability of these platforms to the spread of rumors and misinformation [19,21,38].

Online Rumoring during Crisis Events

The dangerous potential of online rumors, especially in the context of crisis events, is a popular refrain [26]. Emergency managers cite the fear of misinformation—and the difficulty in discriminating between true and false information—as a barrier to adopting and utilizing social media in their work [19]. The rumor issue may indeed be directly tied to the affordances of social media—e.g. how they enable extremely rapid information-sharing and re-sharing [20] and how re-sharing and re-mixing can cause tweets to lose context [9], making it hard to identify the provenance and assess the credibility of a specific piece of information. However, rumoring in the context of crisis events is not unique to social media or the Internet.

Rumoring as a Collective Sensemaking Process

Social psychologists have been investigating the dynamics of rumor spread since at least the 1940s [2]. This work has often connected rumors to the crisis context, where information uncertainty and ambiguity along with individuals' anxiety about impacts and potential responses, act as drivers for the development and spread of rumors [2,36]. Shibutani framed the telling of rumors—i.e. rumor*ing*—as a collective sensemaking process whereby people come together and attempt to make sense of imperfect and incomplete information [36]. In this perspective, rumors are not necessarily false, but can be true, false, or somewhere in between. Additionally, rumoring is not an inherently bad activity, nor is it driven primarily by ill-intentions. Aligned with Quarantelli's assertions that the majority of people experiencing disasters are pro-social and active [32], some rumor participants in this context are motivated by the potential of helping others [20]. Others participate as a cathartic activity—to reduce their own anxiety about the events [7,35]. Though researchers have explained rumoring as a natural feature of crises, it can be viewed as bad behavior and consequently there may be a social cost to passing along rumors [34].

Online Rumoring and the Self-Correcting Crowd

Previous studies have explored the online rumoring phenomenon through its digital traces, both from quantitative [31,33,37] as well as mixed-method and qualitative approaches [3,29,38]. Some of these studies look at rumors generally [31], while others focus specifically on the context of political discourse [33] or the spread of rumors during crisis events [11,28,38]. Though much of the research in this space [11,31,38,44] includes a stated goal of developing methods to automatically detect rumors, another common focus is on the causes or motivations of rumor-sharing. In a study based on statistical analysis of manually-coded tweets, Oh et al. explored online rumoring as a process of collective sensemaking and attempted to identify the causes of rumor propagation, showing that unclear information source, personal involvement and anxiety were factors in rumor spread [29]. Tanaka et al. used an experimental study to assess factors related to retransmission of rumors during the aftermath of the 2011 Japan Earthquake [42], finding that users' determination of the information's importance—but not factors such as perceived credibility—were predictive of intention to pass along a rumor tweet. These findings align with other work [2,20] suggesting that a motivation to be helpful to others is a factor in spreading online rumors.

Several researchers have also specifically examined online corrections [11,28,38,44], almost all with the stated aim of rumor detection. Mendoza et al. explore the dynamics of rumors in the crisis context, finding that rumors tended to be questioned or challenged more by the crowd than factual reports [28]. Their work concludes with a hypothesis that rumors could be automatically detected by algorithms using a content-based approach that identifies questioning or challenging tweets. Castillo et al. follow up that work by presenting a machine learning approach for assessing the credibility of tweets [11]. Zhao et al. build upon this idea of detecting rumors early in their lifecycle by identifying what they call "enquiry" tweets—tweets that they define loosely as seeking more information or expressing skepticism about a story [44]. Andrews et al. examine the role of official accounts in propagating and correcting rumors, demonstrating that official corrections could help to slow or stop the spread of a rumor [3].

Underlying much of the research in this area is the notion of the *self-correcting crowd*—a commonly-held perception that the online crowd will identify, challenge, and eventually correct misinformation. Journalists have referred to Twitter as a "self-cleaning oven" [15] and a "truth machine" capable of "savage" corrections [18]. This idea builds upon theories of collective intelligence [23,43] and the popularized notion of "wisdom of crowds" [41], which claim that collections of individuals can exhibit intelligent behavior in their aggregated activities that exceeds the abilities of any single individual. Mendoza et al. invoke this principle when they claim that the Twitter crowd can act like a "collaborative filter" for information [28]. However, subsequent research demonstrates that many false rumors do not get corrected by the crowd—at least not at the rates that they spread [38]. Moreover, few studies explore *how* members of the crowd choose to (or not to) correct online rumors, or the different strategies they use to do this work.

METHODS

This research is primarily based on fifteen interviews that we conducted with Twitter users who participated in the propagation or correction (or both) of two specific rumors that spread during two distinct crisis events. We draw from two distinct rumors/events not to draw sharp comparisons, but to identify convergent themes across rumor participants—i.e. to identify and articulate strategies and motivations of rumor-correcting behaviors that may exist across rumor and event types. This study is primarily qualitative, utilizing a grounded, interpretivist approach to gather and analyze interview data. However, we employ a range of other methods in this research—e.g. log analysis, content analysis, visual analysis, descriptive quantitative analysis—to identify these behaviors and show how they fit into the broader collective activity of online rumoring.

Our analysis of correction behavior on social media during crisis events begins by capturing digital traces on Twitter. These traces are collected by the research team in real-time using the Twitter Streaming API. We then identify specific subsets of tweets that are related to particular rumors from within these larger event-level collections. Next, we manually classify each unique tweet in each subset as to whether it affirms or denies the rumor. Following this, we generate a behavior pattern or 'signature' for each user according to their rumor affirming/denying actions over time. This allows us to identify individuals who demonstrate different patterns of corrective behavior (for instance, a person who switched from passing along lots of messages that spread the rumor, to ones that correct the rumor). Finally, we interview Twitter users who exhibited different patterns to better understand their motivations and rationales for the rumoring and correcting actions that they took. We unpack each of these steps below.

Step 1: Event Collections and Rumor Scoping

We focus on two false rumors from two crisis events. For both, we captured data using the Twitter Streaming API, executing forward-in-time collections based on keyword search terms selected and curated by our research team.

Case 1: Rumored Hijacking of WestJet Flight 2514

The first rumor is the rumored hijacking of WestJet flight 2514 during the afternoon of January 10, 2015. As the flight was approaching its destination, a flight-tracking site reported that the plane had broadcast a code indicating a hijacking. This rumor soon began to spread on Twitter, as users speculated about the implications of this report—whether or not the plane had indeed been hijacked and, if so, who the culprits were. Eventually, several official accounts including WestJet's became involved in the conversations. In the end, the aircraft arrived in Puerto Vallarta as scheduled and without incident.

Data collection began approximately 20 minutes after the first tweet, at 4:33pm MST on January 10 and stopped at 2pm the next day. We tracked the following terms: *westjet*, *#WS2154*, *hijack*, *hijacked*, and *hijacking*. After ending the collection, to reduce noise from the "hijacking" terms, we scoped the rumor to only include tweets that contained at least one other term related to the WestJet event. The rumor-related dataset for the WestJet Hijacking contains 18,506 total tweets. It is limited by the 20-minute delay in initiating the collection, but we did not experience rate-limiting or other data loss during this event.

Case 2: A Shooting at Les Halles during the Paris Attacks

On November 13, 2015, a series of coordinated terrorist attacks took place in Paris and its nearby suburb, Saint-Denis. At 21:20 CET, three suicide bombers struck near the Stade de France in Saint-Denis, after which suicide bombings and mass shootings took place at cafés, restaurants and the Bataclan Theatre. As these events developed and people attempted to make sense of imperfect and often conflicting information, several rumors began to spread. One erroneously identified the Forum des Halles, a commercial center and an iconic Paris location, as an affected location, claiming that there was a shooting there.

We began collecting tweets at 22:37 CET on November 13, more than an hour after the first attacks, and collected more than 10 million tweets over the next five days. The search term list includes dozens of different terms, including *Paris, ParisAttack, ParisAttacks,* and several other terms related to specific locations that were affected or rumored to be affected—e.g. *Bataclan, Stade de France,* and *Les Halles*.

As the event unfolded, researchers identified Les Halles as (first) a potentially affected site and (later) a false rumor. Once the collection was complete, we scoped the rumor to include only tweets that contained the term "Halles". This resulted in 36,505 tweets. Importantly, the Les Halles rumor did not begin to spread until after our event collection was initiated. However, we did experience substantial rate-limiting during its propagation window and several short periods (~1 minute) of data loss. In this paper, which focuses on interview data, we attempt to report within the constraints of these limitations.

Step 2: Categorizing Tweets

We then manually classify every tweet in the rumor subsets as one of five mutually exclusive codes: Affirm, Deny, Neutral, Unrelated, and Uncodable. We code tweets that support or pass along the rumor as Affirm, and tweets that correct or refute a rumor as Deny. The Neutral category is assigned to tweets that relate to the rumor, but do not take a stance on it. Tweets are labeled Uncodable if they contain words that cannot be deciphered by the researchers, including any non-English words. Significant for the research here, we include only English-speaking tweets in our analysis. Prior work describes this coding scheme and process in greater detail [3,4,38].

Step 3: Identifying Deletions

After the manual coding process, we then identify tweets that have likely been deleted by their author. Using the Twitter Search API, we execute a "status lookup" for the Tweet ID of each tweet in the rumor subset. If that lookup does not return a tweet, then we label the tweet as likely deleted. (Other reasons for it to be missing include a change in a user's privacy settings or account suspension.) For deleted tweets that are retweets, we attempt to determine (by looking up the original) if the deletion was made by the retweeting user or by an upstream author. For the WestJet rumor, deletion identification occurred ten weeks after the event. For the Les Halles rumor, with the goal of moving quickly to the interview phase, deletion identification occurred four days after the event.

User Group	Behavior	Interviewed
Affirm-only	Users post one or more tweets affirming the rumor.	LH9
Deny-only	Users post one or more tweets denying the rumor.	WJ3, LH11
Affirm-Deny	Users post one or more tweets affirming the rumor and one or more tweets denying the rumor.	WJ1, WJ5, LH1, LH2, LH5, LH6, LH7, LH10
Affirm-Delete	Users post one or more tweets affirming the rumor and then delete one of those tweets.	
Affirm-Delete-Deny	Users post one or more tweets affirming the rumor, one or more tweets denying the rumor, and deleted at least one tweet.	WJ2, LH3, LH4, LH8

Table 1. Descriptions of User Groups and List of Participants WJ = WestJet Interviewee; LH = Les Hall Interviewee

Step 4: Generating User Behavior Signatures

Next, we construct a *user behavior signature* for every user who shared a rumor-related tweet in one of the rumor subsets. We log three kinds of user actions relevant to corrective behavior: affirms, denies, and deletions. We use this log to create a behavior signature for each user that sequentially summarizes their recorded involvement in the rumor. For instance, a user who posted two tweets affirming the rumor, deleted them, and then added a tweet denying the rumor would have the signature "A(Del) A(Del) D" and be assigned to the Affirm+ Delete+ Deny+ Group. Table 1 offers an overview of these different user groups with the number of interviewees from each group.

Step 5: Interviewing Diverse Rumor Participants

To better understand how Twitter users experience and reflect upon their rumoring and rumor-correcting behaviors, we conducted interviews with people who had participated in one of these rumors. To gain insight into different kinds of user behaviors, we attempted to interview individuals with different types of user behavior signatures. The signatures therefore served as a mechanism for enhancing the diversity of our interview sample.

Interview Recruitment

For recruiting, we selected users from each user behavior group and reached out to them through Twitter. The selection process was random, though we removed selected users from the pool if their account profiles or recent tweets contained abusive or profane content. The initial contact tweet (which was public) was vague—i.e. we did not specifically mention rumoring—and requested follow-up communication through a private channel (DM/email).

Not surprisingly, the overall response rate was low. Of 185 total recruitment attempts, we interviewed fifteen participants. For the WestJet rumor, interviews occurred between three and four months after the event. For the Les Halles rumor, interviews were conducted between five and eight weeks after the event. There was also a significant selection bias in the respondents: those from the Affirm-only and Affirm-Delete groups were far less likely to respond to our interview requests.

Interview Protocol

We completed 15 interviews total (4 from WestJet and 11 from Les Halles). Of these, 8 were men and 7 were women. Surprisingly, though perhaps related to the self-selection bias, five self-identified as journalists. For the Les Halles rumor, four participants lived in Paris—and two were actually living near Les Halles at the time of the attacks.

Interview Protocol

We conducted in-depth, one-hour interviews with each participant. All were remote, conducted via Skype or phone. All but one were in English. For the final interview, LH11, a native French-speaking member of our research team interviewed the participant in French. Except for LH11, each interview was carried out by two researchers (with a third usually acting as a silent note-taker). All interviews were recorded and transcribed.

We used a semi-structured protocol designed to elicit participants' perspectives on their own corrective behavior. Participants were first asked about how they learned about the event, their impressions of the information space at the time, their motivations for participating in information sharing on Twitter and their intended audience (if any). Following this, we asked questions designed to help them speculate and reflect upon what kinds of things they would do in future events (and why) if they realized they had posted misinformation. At the midpoint of the interview, as a cue for more specific questions, we provided participants with a log of the tweets we had collected and used to identify them for recruitment. We then asked them to talk us through these tweets, to explain their motivations and intentions. Several also chose to browse through their social media history as a memory aid during the interview.

Step 6: Interview Data Analysis

In analyzing the interview data, we took a grounded, inductive approach to help us organize and surface patterns from the data. As a first step, our research team (eight individuals) carried out an open-card sort on the interview transcripts. Each was atomized into individual statements and printed onto a card, then the research team clustered these cards based on similarities. This clustering process yielded multiple emergent categories that were iteratively merged, removed or split based on the research team's perceptions of their usefulness, descriptive power and scope. After discussing and refining these categories, we settled on a small set of salient themes that seemed most relevant to our initial questions about corrective behaviors.

In a subsequent phase of focused-coding, we returned to the original transcripts to identify content related to our refined list of themes. For each thematic category, two researchers went through each transcript looking to identify each instance of that theme. Researchers also generated memos, articulating their ideas for how interview content connected to specific themes and how themes connected to each other. Additional sub-themes emerged during this phase as well.

Ethical Considerations: Identifying Deleted Tweets and Interviewing Rumor Tweeters

We encountered significant ethical challenges around working with deleted tweets as well as recruiting, interviewing and reporting results from online users who participated in a socially stigmatized activity: passing along rumors. For the deleted tweets, we followed the protocol outlined in [25]—removing from content analysis any tweet that had been deleted once we identified it as a deletion. We made an exception for the deleted tweets from the interviewees, who consented to participate in this study. We also attempted to reduce stigma during the interviews themselves by telling participants that rumors are a natural part of disaster events and, for the Les Halles rumor, that one of our researchers had also shared a rumor-affirming tweet. To reduce the risk of damage to participants' reputations, we have anonymized all usernames, changed some demographic data, and have not included any actual tweets in our reporting. Where we refer to tweet content, we have significantly altered the syntax and structure of the original tweet along with other details like the time, number of retweets, etc. to prevent discovery of its original author.

TRACE ANALYSIS FINDINGS

Rumor 1: WestJet Hijacking

The WestJet rumor began with a notification on a flight-tracking website that Flight 2514 was "squawking" code 7500, the code for hijacking. Shortly thereafter, a user took a screenshot of that report and posted it to Twitter. A rumor that the flight had indeed been hijacked quickly began to propagate, as aviation fans and breaking news accounts spread the information to an increasingly wide audience. Peak volume exceeded 400 tweets per minute about thirty minutes after the initial report.

Code Category	Total Tweets	# Deleted	% Deleted	# Actively Deleted
Total	21,057	3662	17.4%	2651
Affirms	8438	2165	25.6%	1394
Denies	8064	852	10.6%	722
Neutral	1013	148	14.6%	140
Uncodable	2551	387	15.2%	299
Unrelated	991	110	11.1%	96

Table 2. Deletions by Tweet Type for WestJet

Unlike the majority of Twitter rumors where affirming tweets dominate [e.g. 38], in the WestJet rumor there are almost as many Denies as Affirms (46% of related tweets vs 48%). Most striking is a dramatic shift from mostly affirming tweets to mostly denying tweets. This shift occurred about 45 minutes into the rumor's lifecycle and immediately after WestJet's official Twitter account began to post rumor-denying tweets—official corrections that were widely retweeted. Previous research suggests that the both the shift and the relatively high volume of denials in this case were related to efforts by the official WestJet account to refute the rumor [3].

Tweet deletions are another interesting feature here. Recent research suggests ~11% of tweets are deleted [6]. Aligning closely with that number, when we captured deletion information (two months after the event), 11.1% of Unrelated tweets and 10.6% of Deny tweets were missing. However, 25.6% of Affirm tweets (nearly twice the baseline rate) were missing, and 64% of those appeared to be "active" deletions—i.e. not the result of deletion cascades. This lends evidence to previous claims [25] that for false rumors, affirming tweets are more likely to be deleted than denying tweets.

In our dataset, there are 8963 users who shared a tweet related to the WestJet rumor (Table 3). The two largest groups of users exhibited the Affirm-only pattern (39%) and the Deny-only pattern (28%).

About one-third of rumor participants posted tweets demonstrating a shift from one rumor stance to another (e.g. from affirming to denying). Of these, 1728 users (19% of the total) sent at least one Affirm and one Deny with no deletions, 806 (9%) users sent one or more Affirms and actively deleted at least one of them, and 406 (4.5%) sent at least one Affirm, deleted at least one tweet, and also posted at least one Deny. Deletion was a prominent behavior in this rumor—more than 10% of users who participated in this Twitter rumor deleted at least one of their tweets.

User Behavior Signature	Total Accounts for WestJet	Total accounts for Les Halles
Total	8963	4589
Affirm-only	3476	4097
Affirm-Deny	1728	13
Affirm-Del	806	283
Affirm-Del-Deny	406	3
Deny-only	2547	193

Table 3. Accounts by User Behavior Signature

Rumor 2: Les Halles Shooting during the Paris Attacks

On November 13, 2015, a series of terrorist attacks took place in and around Paris. Though the attacks were coordinated, they took place at different times and some, like the siege at the Bataclan Theater, lasted for extended periods of time. As events developed, Twitter users responded with first-person accounts, attempts at providing material and social support, and information about the location of the attacks. The individuals we interviewed, who include both locals and remote onlookers, described this time period as one of high uncertainty and anxiety:

LH7: "It's really hard to convey how little everyone knew. There were so many rumors flying around. Where there were attacks going on. How many attacks there were. Whether or not they were coordinated. Whether or not it was all a hoax or prank. No one knew anything for sure. The only thing people knew anything of at first was the thing that happened at the football stadium. But all the other little things happening in the different parts were kind of hearsay at first… things going around on Twitter about Les Halles, about the Louvre, about so many places in Paris that weren't at all targets as it turned out."

The rumor about Les Halles began at approximately 11 p.m. CET with a French language tweet stating that there was a shooting there. That tweet was highly retweeted, and was followed by a wave of similar messages claiming an attack at that site. Several mainstream media sources helped to spread the rumor through their broadcasts, websites and social media accounts. Europe 1 radio was an early source. France 24, BBC, SkyNews, Reuters, FoxNews and others also helped to spread the rumor in some capacity. The rumor peaked on Twitter about one hour after it began and then experienced a period of exponential decay, functionally disappearing about twelve hours later.

Overall, the denial signature for this rumor is weak—only 4.4% of related tweets were denials. Though small in relative volume (against affirms), the rate of Denies surged slightly just after 12am CET, due to tweets from (and retweets of) a few individuals on the ground near Les Halles. But despite these first-hand denial tweets and the activity around them, the denial rate never surpassed the Affirm rate (as we saw in WestJet). This denying activity does seem to correspond with a drop in the rate of affirming tweets—the tweet rate lost more than 50% of its volume during a 20-minute window when the denial surged. Research [38] suggests that such patterns in volume—where affirms show a dramatic spike and then slowly fade away, and where denials never reach the same peak tweet rate as affirms—are common for rumors during crisis events. In this sense, the Les Halles rumor can be considered more 'typical' than the WestJet rumor.

Code Category	Total Tweets	# Deleted	% Deleted	# Actively Deleted
Total	36,505	4131	11.3%	2792
Affirms	4790	621	12.9%	308
Denies	224	6	2.7%	6
Neutral	37	3	8.1%	3
Uncodable	22,177	2296	10.4%	1302
Unrelated	9277	1205	13.0%	1173

Table 4. Deletions by Tweet Type for the Les Halles

We captured deletion information for this rumor four days after the event. At that time, 13% of affirms were missing and 50% of those appeared to be "active" deletions. Conversely, only 2.6% of denies were missing and all six of those were considered to be active deletions. Though these data are not directly comparable with the WestJet data due to a shorter gap between the event date and the deletion identification, they reinforce the claim that Deny tweets are less likely to be deleted than Affirm tweets.

Table 3 provides a breakdown of 4589 users we identified having shared a tweet related to the Les Halles rumor. The vast majority only sent Affirm tweets, with 89% of users in the Affirm-only group and 6.2% in the Affirm-Delete group. Less than 5% of accounts sent a Deny tweet, and most of those were in the Deny-only group. We only identified thirteen users in the Affirm-Deny and three users in the Affirm-Delete-Deny groups. Interestingly, users who affirmed the rumor were much more likely to take the correcting action of deleting a tweet than sending a denial.

INTERVIEW FINDINGS

Corrective Objectives

Our analysis of interviews with Twitter users who participated in rumoring and correcting rumors uncovered three different objectives for taking correcting actions: *correcting oneself*, *correcting the information space*, and *correcting another person* (or organization). For each objective, there are different types of actions that can be taken—for example, in correcting oneself, a user can choose to delete a rumor-affirming tweet or to post a correction. Our interviews revealed that, even within a single rumor, some Twitter users employed multiple types of corrective actions and often considered others that they elected not to use. Using the user behavior patterns as an initial guide, and the interviews to unpack those patterns, in this section we describe users' rationale for their corrective actions in relation to the three correction objectives.

Correcting Oneself

The first, and perhaps most obvious objective for taking corrective action related to Twitter rumoring is to correct oneself. In this case, a user has posted a rumor-affirming tweet and later becomes aware that the information they shared is either untrue or unconfirmed.

Affirm-Deny: One action a user can take to correct herself is to post a Deny tweet. One way to do this is an *explicit self-correction*, where the user specifically addresses the fact that she shared a rumor-affirming tweet. When asked about their general strategies for correcting themselves after passing along a rumor, four participants stated they would post a follow-up tweet to explicitly acknowledge their error and apologize. However, these explicit corrections are rare in the rumoring data we collected. Far more common among the users we interviewed were *implicit self-corrections*, where after sharing one or more rumor-affirming tweets, a user sends a subsequent tweet that contains information either 1) directly questioning or noting uncertainty regarding information in the original tweet; or 2) implicitly contradicting information in the original tweet.

*Affirm-Delete***:** Another action a user can take after she realizes she posted a rumor-affirming tweet is to delete. A deletion removes a tweet from that user's history and the public timeline. Her followers are no longer able to see it and it will no longer appear in searches. The deletion action therefore can function as a correction of the information space (discussed below) or a self-correcting action.

As a self-correcting action, a deletion without a follow-up correction was considered by some of our interviewees as a form of hiding one's error. Seven shared negative opinions about this strategy, and though we attempted to recruit 52 users with this behavior pattern, only one responded to an interview request, and his recollection of his tweeting patterns suggest he may have had a different pattern.

Affirm-Delete-Deny: Some users choose a two-part correction strategy that involves both deleting the rumor-affirming tweet, and posting a denial tweet. Among our interviewees, four demonstrated this corrective action sequence. Two, including LH4, were self-identified journalists who tweeted during the Paris Attacks.

LH4 was a high volume tweeter during the event. The account was actually operated by multiple people constituting a "new media" organization. This account posted two tweets related to the Les Halles rumor, early in its lifecycle. The first was an Affirm, the second a Deny. Both expressed uncertainty, calling attention to the ambiguity around this rumor. Then, about 45 minutes later, the account posted this tweet, clearly affirming the rumor and providing (false) evidence:

```
Photo from the shooting at Les Halles in
Paris. <URL>
```

According to the account operator we interviewed, within ten minutes, she deleted that tweet and posted a correction:

```
That image was not from Les Halles. Our
previous tweet has been deleted. Sorry.
```

This corrective action sequence functions to 1) remove the false information from the broader information space; and 2) to draw attention to the fact that the information has been challenged or corrected. Interestingly, this denial tweet is the only explicit self-correction shared by any of the interviewees in this study. Though there are occasions where it might not be ideal, this action sequence can be considered both altruistic (in terms of removing false information) and honest/transparent (as the user openly admits to her mistake). We return to and build upon those distinctions in a subsequent section of this paper.

Affirm (only): Finally, users can choose to take no action. By far, the most common "correcting action" across our data set (for those who sent a rumor-affirming tweet) was no action. 54% of users who shared the WestJet rumor and 93% of users who shared the Les Halles rumor sent only affirming tweets and did not take any direct action to correct them. We cannot make assumptions about how many came to know that the information they shared was false, but our interviews suggest that the absence of corrective action does not mean that someone did not become aware a rumor had been challenged or corrected.

One reason that participants gave for not correcting, especially in the case of the Les Halles rumor, was continued uncertainty about the rumor. Unlike the WestJet rumor, for which there was an official correction within an hour of its origin, the Les Halles rumor did not see such a quick or firm resolution. Six interviewees explained that even after many users, including some "on the ground" in Les Halles, began to tweet denials, they still were not sure about the rumor's veracity. This ambiguity may have discouraged people from correcting. LH1 explained why someone might hesitate or choose not to correct in this situation, "[I would have to be] certain it was a definite false alarm before I would go back and say 'yes it's a false alarm', I think. I would have to be pretty definite."

Other users expressed they did not feel the need to self-correct. LH9 suggested that the burden of false information lies with the consumer. He sent hundreds of tweets (almost all retweets) related to the Paris Attacks. Four of those were affirmations of the Les Halles rumor. He did not delete or correct any of those tweets. He explained that since he was not tweeting this information to anyone in particular, he did not need to correct it. Acknowledging the role of imagined audience, he went on to say that if he had misinformed someone he knew personally, then he would have let them know. In other words, in his view, other people seeing his tweets are responsible for verifying this information for themselves, except for those to whom he has close ties. Implied in this rationale is an argument that downstream users should verify their sources, and that being a close tie is a form of verification.

Correcting the Information Space

A second objective for corrective behavior is to correct the information space. Almost all of the interviewees who took corrective action noted that, on some level, their motivations were not necessarily to correct a previous error they had made, but to make sure the information spreading through Twitter was as accurate as it could be. LH2 summed up this orientation, "I was concerned with trying to not allow rumors to spread. I wanted to make a modest contribution in which to clarify what was happening and not allow rumors and misinformation." Four correcting action sequences were associated with this objective.

Affirm-Deny: Several interviewees who exhibited the Affirm-Deny pattern explained their objective as correcting the information space, not themselves. LH7 is an interesting case. She was living near Les Halles during the Paris Attacks. Initially, she was gathering information through Twitter. At around 22:40 UTC, she saw tweets about the Les Halles rumor. She retweeted one of those tweets and then shared her own original tweet stating that there was an attack at Les Halles. Then she left her apartment and went out to verify for herself if that rumor was true. About ten minutes later, she shared two tweets similar to:

> I'm in Les Halles. People don't understand the reports of a shooting in the area. Police on the scene have left. #ParisAttacks

Her rationale for the rumor-denying tweets was not related to her previous affirming tweets. Instead, it "...was to correct the information that was out there. In the perspective of someone who is there, rather than the hearsay that was going around Twitter." Due to her location on the scene, she recognized that she had 'ground truth' information to share, and she wanted to use that position to get the best information out.

Affirm-Delete & Affirm-Delete-Deny: Some users also explained deletions of rumor-affirming tweets as a way of improving the information space—i.e. by deleting the tweet, the user takes it out of the public stream. People will no longer be able to see or retweet that post, and previous retweets of the original will be automatically deleted as well. The act of deleting can therefore be viewed as one of trying to stop a rumor from spreading.

LH4, the "new media" account whose rumor behavior is featured in the self-correcting section above, described a nuanced rationale for deletions, and related those directly to the potential impact the rumor-affirming tweet would have on the information space:

"We very, very, very rarely delete tweets. It's pretty much never will we do that, unless we are so worried about incorrect information getting out that we have to delete it. So this was one of those case where we were like, 'I think we need to delete this because it's gone really completely blatantly wrong and it's going to be retweeted a lot because everything we [are] doing [is] getting retweeted a lot.' So we decided to delete it."

Similarly, WJ2 described the rationale for deleting in this way, "we found out what actually happened and I was like, alright there is no point in putting false information out there, no point in having everyone see it and come to conclusions. So I got the truth, and I am going to delete the false information that is not real because no need for other people to see it."

Deny (only): Another user pattern in our data is one of only denial tweets. 28% of users in the WestJet rumor and 4% of users in the Les Halles rumor only posted Deny tweets (according to our data collection and coding). In these cases, a user is clearly not correcting themselves. Instead, they are often trying to contribute to a better information space. Previous research has identified information verification (and rumor challenging) as a core task taken on by digital volunteers during crisis events [39].

When asked about this behavior, WJ3, who sent four denial tweets of the WestJet rumor, positioned his actions as targeting the information space, not a specific individual.

WJ3: *"No I was not correcting anyone, I was just providing information…. I was not taking aim at the people who were speculating. I was trying to find the best, correct information that I could."*

WJ3 was not specifically correcting another person, but simply trying to share the best available information at the time. Other interviewees (with other user behavior patterns) shared similar sentiments about not directly correcting or challenging other users about their rumor-affirming tweets.

Correcting Another User

Direct, explicit corrections of other users were rare in our data set. One rough measure of this behavior is to identify denial tweets that begin with a "@" or a ".@". This results in only 34 tweets within the WestJet rumor and 10 tweets in the Les Halles rumor. After our first round of interviews, noting an important gap in our participant pool related to this behavior, we purposefully recruited and interviewed a user who shared a Deny tweet in this form, LH11. LH11's one denial tweet was addressed to a breaking news account:

> .@<breaking new account> it isn't at les halles, it's at the bataclan, stade de France and the place de la République"

LH11 reported that he had been confident in posting this correction because he had firm evidence that this rumor was false—from calling a relative who was at Les Halles. LH11 held to the rationale that correcting others and oneself is the right thing to do if it means 'redirecting people' towards the right information.

To Delete or Not to Delete

An important consideration for users who realize that they shared a tweet related to a false rumor is whether to delete that tweet or not. Interviewees described extremely nuanced heuristics for this decision, noting a number of different conditions and factors. One theme that emerged repeatedly was the tension between maintaining an accurate personal tweet history (a reputation concern) and maintaining an accurate information space (an altruistic concern).

Multiple interviewees expressed a reluctance to alter the historical record by deleting their tweets. Their explanations suggested two related concerns: 1) deleting a tweet is a potential method for protecting reputation by hiding an error; and 2) this kind of correcting action is perceived as deceptive and dishonest. LH4 described how her colleagues and she tried to be "100% transparent*"* and that they "very rarely delete tweets". WJ1 explained that "...you don't want to make it seem like you are deleting your record of what you tweeted previously." And LH6 stated, "As a general policy I am not revising my history."

However, interviewees were also aware that there were cases, especially around rumoring tweets in this context of crisis events, where deleting a tweet might be a better—or more altruistic—strategy.

WJ3: *"There are some people who say you should never delete. I am not that type of person. For some, deleting might be the best practice. The worst thing is to have bad information out there, particularly on Twitter, because any individual tweet not seen in a stream is 'out of context.' If that tweet is seen and you do not see the correction tweet, then that is true."*

In the above quote, WJ3 explains that there is danger in not deleting. Because of the way information propagates on Twitter (and other online spaces), simply sending a follow-up denial tweet may not stop a rumor from spreading. If the original affirming tweet is not deleted, it can continue to propagate, e.g. as retweets, and those downstream tweets may not retain a connection to the denial.

Interviewees described a detailed set of criteria upon which they based the deletion decision. One issue was timing. If the author thought the tweet had been out there for a while, they would not delete it. But if they felt it had just recently been sent, and that few people had seen it, they would delete it. A second concern was interactions—i.e. how many retweets or mentions their rumor-affirming tweet received. At first glance, interviewees offered seemingly contradictory heuristics here. Several rationalized that if a rumor-affirming tweet had received a lot of retweets or if it might receive a lot of retweets, then it should be deleted. These "active" tweets, as one participant termed them, are beneficial to delete because retweets will be deleted as well. LH6 expounded, "If I tweeted something that turned out to be erroneous, and it had 40-50 retweets or there is a lot of action happening on it, it would make more sense probably to delete the tweet." LH4 explained that the potential of being retweeted at a high volume, even if it was not yet happening, was a good rationale for deleting a rumor-affirming tweet. On the other hand, LH7 said that if a rumor-affirming tweet had several retweets, it would be a reason *not* to delete it. LH8 took a more relative approach, noting the deletion cascade effect and drawing a distinction between retweets and other kinds of interactions:

LH8: *"If it's retweeted and I delete it, I think it's deleted from all the other feeds. So I would definitely delete it. And if there is a conversation that's a bit tricky [...] If there is a conversation I don't know if I would delete it because sometimes you come to a conclusion that the information was partially wrong or partially right."*

A final consideration that was shared by multiple interviewees involved the weight of the error—i.e. how much damage it might do. Three interviewees told us that minor errors, like typos, were okay to delete. At the other end of that spectrum, tweets that could cause major harm, for example to another person's reputation, were also considered okay to delete.

Considered together, these diverse rationales for whether or not to delete a tweet demonstrate how the imagined audience [8,24], including not just who that audience is and how it will perceive the user, but also how that audience will act upon the information, contributes to a rumor-tweeter's decision on whether and how to correct.

Locus of Responsibility

When asked to review and explain their tweeting actions, almost all participants provided rationalizations for why they were not fully responsible for having shared the rumor. Though likely a natural response to the question format, these rationalizations shed light on users' perceptions for how online rumoring takes place and what roles they play in this rumoring. During the initial card sort of our interview data, we noted the salience and diversity of these rationalizations and how they connected to different kinds of behaviors. We identified six distinct—though in places overlapping—perspectives on the "locus of responsibility" for sharing and correcting rumors.

Self: Taking Responsibility for Sharing/Correcting Rumors

Some participants expressed a sense of personal responsibility for posting the rumors—blaming themselves for not having verified adequately and noting concern about how their posts may have affected others.

LH4: *"This one was me and I was wrong. I had neglected to notice [how this information was not related to Les Halles]. My [colleague] came back online and said 'Hey, we already tweeted that earlier today. It's [not Les Halles]. I've verified that.' I'm like 'Oh my gosh, I'm so sorry.'"*

As LH4's quote shows, when a person draws the locus of responsibility inward, it can be uncomfortable. They may feel ashamed for what they view as an error—and a public one. They may also feel responsible for causing others to see and share the rumor. At least two other interviewees expressed significant distress about their rumoring activity.

WJ1: *"I was left with a sense of anxiety after the whole thing was over, and I feared I had needlessly alarmed or frightened people, and I worry about that event. I feel uneasy about it. I am not sure it is possible to avoid this kind of feeling, but it left me with a feeling of unease."*

WJ1 describes his anxiety as being rooted in a sense of responsibility to those who may have been misinformed by his tweet. Here he positions the relevant "downstream" individuals as people who may have had a loved one on the flight, but this sense of responsibility was also seen to apply to other Twitter users who simply read and passed along his tweet—those had been drawn into the rumor by his tweets. This demonstrates the role of imagined audience—in this case a concern for how members of ones' audiences will perceive ones' actions, as well as how those audiences will be affected by those actions.

Upstream User: "But I Heard it From <username>"

In many cases, participants were seen to place the locus of responsibility on the source of the information, whether a trusted news source or a friend. For example, LH3 had lived in Paris and had worked with one of the mainstream news outlets there. In explaining why she sent rumoring-affirming tweets about Les Halles, she stated that both were retweets of major news sources whom she trusted. LH5 also noted the role of mainstream news outlets in his rationale:

"[I tweeted] because of this person citing this. ... I probably should have been vague, but the fact of the matter is that when you see <news outlet> reporting it, I'm like 'OK'. It seemed much more real."

Like LH3 and LH5, interviewees who invoked this perspective deflecting responsibility to the source typically pointed the finger at trusted media outlets. This sentiment, which was shared to some extent by the journalists we interviewed, suggests that online rumor participants have different expectations for journalists compared to other members of the crowd. However, users assigning blame to upstream sources also directed that blame at trusted friends and other accounts they were following.

Downstream Users: "They Should Have Verified My Tweet"

A small set of interviewees placed the locus of responsibility on downstream users—their followers and others who might be reading and re-posting their tweets. In their interview responses, they rationalized their rumor-sharing behavior by suggesting that their audience should not be accepting their tweets as fact, but should be verifying this information themselves.

Crisis Events: "That's Just the Nature of These Events"

Other interviewees emphasized that rumoring is just a natural part of crisis events. They described their motivations as trying to help other people by getting information out quickly. They noted the uncertainty and ambiguity in the information space and how difficult it is to discern good information from bad, truth from rumor. Two participants highlighted a difficult trade-off: is it worse to pass along this rumor (in the case that it turns out to be false), or to not pass along this rumor (in the case that it turns out to be true)? Often, it seems, the default answer is tweet now and worry later.

Locus of Responsibility	Description
Self	Takes responsibility. Likely to correct.
Downstream Users	Says people should verify info themselves.
Upstream Users	Says they trusted source who got it wrong.
Crisis Events	Just the nature of crisis events.
Twitter	Affordances of the platform lead to rumors.
Crowd	Says the "crowd" should/will correct.

Table 5. Locus of Responsibility Categories

Twitter: "That's Just the Nature of Social Media"

Several interviews pushed some responsibility onto the platform mediating the rumor. There was a common perception across almost all of the interviewees that the real-time nature of Twitter was both a huge advantage for it as a place to seek information during disaster and a major contributor to the spread of false rumors. For some, this awareness was a cautionary point, something that one should take into account as they participate. But for others, this perspective could lead to more of an acceptance that these are just the limitations of Twitter, that you cannot expect to it be what it is not. LH8 touches on this:

LH8: *"I think most people, like me, they trust TV more than Twitter because when you're on Twitter you know that people post things that they have not checked before. That's why being a journalist is a job, because you checked your sources first, and this is not the case for Twitter. But when you know this, it is fine."*

For some interviewees, this attitude included a hint of resignation and an abdication of responsibility. However, for the journalists in our set, this came with a new set of competing responsibilities. All were quite reflective about the challenges of balancing journalistic expectations with the pressures to of keeping up with and staying relevant within real-time news.

The Crowd: "The Crowd Will Fix It"

Finally, several participants provided explanations placing the locus of responsibility for rumoring and rumor-correction on "the crowd." Some expressed an implicit trust in the crowd, using it to help verify information, for example through triangulation or by waiting to see if a story "has legs." Considering rumor corrections specifically, several interviewees expressed a more explicit trust in the crowd when it came to identifying and correcting rumors:

LH7: *"I think [rumoring is] part of Twitter and something we have to understand ... that's going to happen. It's like a* *free information sharing tool. Everyone has freedom of speech (hopefully) and hopefully if someone is spreading false information, that information is quickly debunked through other people responding and giving their own insight to something."*

Comments such as this can be linked back to the notion of the self-correcting crowd—i.e. that the online crowd will naturally identify, challenge and ultimately correct misinformation propagating among its members. This idea, which has been popularized in the press [15,18], can be viewed as a kind of "folk theory" [14,22] of how social media systems function.

DISCUSSION

Through the analysis of interviews and Twitter data related to two rumors in two significantly different crisis situations, this paper illustrates how online users engage in rumor-correcting behavior. In this section, we first synthesize the three components of the findings into a preliminary model of rumor correcting. Finally, we explore how imagined audiences [8,24,27] and the broader concept of folk theories [14,22] guide the actions users take to correct online rumors.

A Model of Decision Making for Crisis Rumor Correction

After encountering conflicting information about the veracity of an online rumor, the decision of whether and how to correct has multiple inter-related factors. This paper identifies and explores three components of this decision-making process (see Figure 1).

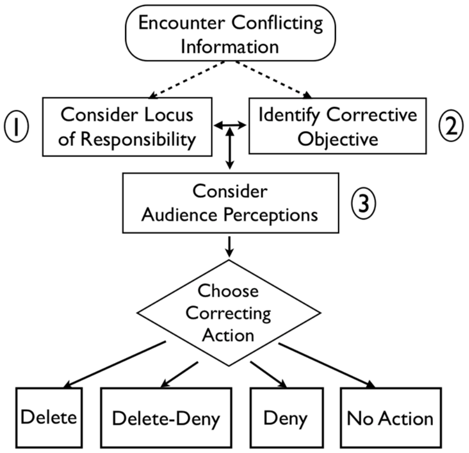
Encounter Conflicting Information
①
Consider Locus of Responsibility
Identify Corrective Objective
②
Consider Audience Perceptions
③
Choose Correcting Action
Delete
Delete-Deny
Deny
No Action

Figure 1. Decision Making for Twitter Rumor Correction

One salient component is the *locus of responsibility*. This includes consideration of who is to blame for the spread of false rumors as well as who has the power to correct them. It also includes how a user conceives of her role within that group. For example, if a user places responsibility in the crowd (espousing a belief in the "self-correcting crowd"), does she see herself as part of that crowd and as capable of playing a role in the correction? If she considers journalists to have a different set of obligations regarding rumor correction, does she see herself as a journalist? If so, then she may take different actions than someone who does not.

A second component is to identify the *corrective objective*—i.e. oneself, another user, or the information space. This consideration is shaped by where the user places the locus of responsibility as well as whether or not that user shared a rumor-affirming tweet. For example, a user who positions himself as the locus of responsibility and has shared a false rumor (e.g. WJ1, LH4) will often choose to correct himself. The corrective objective can also act to shape considerations of the locus of responsibility—for instance, LH7 provided a rationale suggesting her first order concern was to correct the information space and that she later came to realize the significance of her position within that space—as a local authority [40]—and herself at the locus of responsibility.

A third component is to consider whom one's audiences are—i.e. the *imagined audiences* [8,24,27]. This is not just limited to conceptualizing the people with whom we are communicating and their immediate reactions, but also how they will act upon the information we share with them. For instance, one concern that arose repeatedly in the interviews was a perceived trade-off between accuracy and transparency: Deleting can be perceived as a sign of inauthenticity, but is it better to leave it out there where it may mislead others? Another concern was the social impact of explicitly correcting another user—something many users tried to avoid. Though we noted that this was primarily a downstream concern for our participants—likely due to the way that interviewees (and perhaps all users) rationalize their behavior after the fact—it is likely that reputational concerns have a shaping effect on corrective objectives and the locus of responsibility. For example, the perceived reputational impact of an explicit self-correction might cause an individual to revise his corrective objective (e.g. from himself to another) and subsequently reassess his position on locus of responsibility (e.g. from himself to an upstream source).

Finally, users take corrective action. For example, a user who positions herself as at least partially responsible for the spread of a false rumor and chooses to correct herself might post a denial in the form of an explicit self-correction; a user who thinks the crowd will correct and accepts some agency for himself as a crowd-member might choose to correct the information space by deleting his tweet (if he believes accuracy trumps transparency) or by posting an implicit denial (if transparency is more important). Some users choose to take no action. The decision-making process is shaped at each level by whether or not you affirmed the rumor—there are different options (at each level) for those who affirmed and those who did not.

This preliminary model, which emerged from our grounded analysis of interview data, aligns closely with Litt's model demonstrating the relationship between imagined audiences and online action [24]. This sets the stage for a discussion about how users' conceptions of their audiences—as well as their understandings of the broader dynamics of social media—play a role in shaping their corrective behavior.

How Imagined Audiences Shape Corrective Behavior

Research suggests the importance of imagined or perceived audiences in shaping a social media user's actions [8,24,27]. When asked directly, most participants in our study did not acknowledge reputational concerns or attending to their audiences. However, in explaining their rationale for taking certain actions (and not taking others) during the crisis event, interviewees revealed underlying awareness and concerns about their various audiences. A few (LH4 and LH7) expressed sentiments suggesting they were acutely aware of a growing audience, due to their position of influence within the information stream around the event, and several interviewees talked about the perceived trade-off between maintaining an accurate information space (an altruistic concern) and being perceived as trying to hide something (a reputational concern). This evidence suggests that many online rumor-participants are indeed aware of their audiences, and adjusting their behavior according to what they see as the expectations of that audience.

Building upon Giddens' theory of structuration [17], Litt presents a model demonstrating how imagined audiences—which emerge from users' interactions with the system—shape online behavior. Indeed, we see evidence of the shaping role of imagined audiences in our data [24]. Superficially, we can see in our model how considerations of who one's audience is and how they will perceive one's actions play a role in guiding the choice of if and how to correct a Twitter rumor (Figure 1, #3). However, we hypothesize that conceptions of imagined audiences is more complex than that.

Discussions of imagined audience have often focused on a user's conceptions of who an audience is and how that audience might perceive them through their online actions [e.g. 8,24,27]. This kind of dynamic shows up in our data—for example around the interviewees' rationale for not wanting to be perceived as trying to hide an error through a deletion. However, rationale presented by interview participants demonstrates that people are not only trying to understand the size and make-up of their audience, but are also trying to piece together *how* their audience is acting, both individually and collectively, upon what they share.

For example, WJ1 worried about "potentially causing [others] anxiety or making them pay attention to something that they shouldn't have to think about." Another participant noted that he acted the way he did because he was worried that his tweets would create a panic: "…so if you looked at the tweets that I retweeted... I picked and chose as carefully as I could, because while [the situation] was concerning…I want to not suddenly cause a panic. I have enough of a reach that I probably could have caused one, so I was just being very picky about it." Embedded in this concern about panic was an assumption that, under certain conditions, the audience might propagate their messages to a large volume of people. Similarly, considerations about whether and how to correct a rumor-related tweet included theories about how different members of ones' audience might encounter and choose to propagate (or not) the original or the correcting tweet later.

These conceptualizations of not just who an audience is, but how that audience works (in conjunction with system features), are possibly more akin to the "folk theories" that people have about how social media systems function [14,22]. We can see evidence for these folk theories at work within the locus of responsibility categories that emerged in this study (Figure 1, #1). For example, positioning responsibility on the "crowd" reflects the use of the popularized notion of the self-correcting crowd—which takes into account how the "audience" acts upon information and how it reacts to others' actions within the system. Similarly, assigning Twitter (or crisis events) as a locus of responsibility, which many interviewees did at least to some extent, also reflects the impact of folk theories—e.g. about the intersection of technical affordances and human behavior—on the structure that guides decision-making in this context.

This research demonstrates that folk theories guide rumor-correcting actions, and that these folk theories consist of reasoning related not just to how the algorithms work [14], but to how the broader system—including the technological platform or platforms with their affordances, interfaces and algorithms, as well as the other human (and non-human) actors in the system—functions.

LIMITATIONS AND FUTURE WORK

This study has several limitations. The Twitter data we used is incomplete (due to rate limits) and biased (due to the terms we tracked). Though we attempted to utilize those digital traces to identify people who exhibited different kinds of rumor-correcting behavior, self-selection bias shaped the participant sample towards individuals who were more invested in actively correcting the rumor. Additionally, though we asked interviewees to discuss their larger patterns of use across other sites and platforms, the digital trace data, recruiting strategy, and interview protocol render this study highly-focused upon online rumoring within one platform—Twitter. And finally, as with any retrospective study, our interview responses were likely affected by misremembering and post-hoc rationalizations.

However, despite these inherent biases, by connecting actual digital traces to recruitment strategies and interviews, we were able to 1) recruit interviewees who displayed several different correcting behaviors; and 2) provide them with assistance in remembering their actual behaviors (their tweets and deletions). Though some rationalizations (e.g. around why to delete or not) are closely tied to the specific affordances of Twitter and the context of online rumoring, the broader findings about the role of imagined audiences and folk theories of how those audiences interact with data are likely to apply to other platforms and contexts.

The model of rumor correcting presented here is illustrative and functional, but likely incomplete. We introduce it here to synthesize findings, to show how the different constructs fit together, and to provide a foundation for the major theoretical contribution of this paper—demonstrating how the shaping role of imagined audiences in online behavior includes not just who those audiences are but also how they react to and interact with the information we share and the actions we take. Future work may reveal additional considerations and help to further unpack and link together the three constructs presented here.

ACKNOWLEDGMENTS

This research is a collaboration between the emCOMP lab and DataLab at the University of Washington and was supported by NSF Grant 1420255. We also wish to thank the UW SoMe Lab for providing infrastructure support as well as the many students who provided significant assistance to this project.

REFERENCES

1. Adam Acar, and Yuya Muraki. 2011. Twitter for crisis communication: lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities* 7, 3: 392-402.
2. Gordon W. Allport, and Leo Postman. 1946. An analysis of rumor. *Public Opinion Quarterly* 10, 4: 501-517.
3. Cynthia Andrews, Elodie Fichet, Yuwei Ding, Emma S. Spiro, and Kate Starbird. 2016. Keeping up with the tweet-dashians: The impact of 'official' accounts on online rumoring. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16), 452-465.
4. Ahmer Arif, Kelley Shanahan, Fang-Ju Chou, Yoanna Dosouto, Kate Starbird, and Emma S. Spiro. 2016. How Information Snowballs: Exploring the Role of Exposure in Online Rumor Propagation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16). ACM, New York, NY, USA, 466-477.
5. Michael S. Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. 2013. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 21-30.
6. Parantapa Bhattacharya, and Niloy Ganguly. 2016. Characterizing deleted tweets and their authors. In *Tenth International AAAI Conference on Web and Social Media*.
7. Prashant Bordia, and Nicholas DiFonzo. 2004. Problem solving in social interactions on the Internet: Rumor as social cognition. *Social Psychology Quarterly* 67, 1: 33-49.
8. danah boyd. 2007. Why youth (heart) social network sites: The role of networked publics in teenage social life. *MacArthur foundation series on digital learning–Youth, identity, and digital media volume,* 119-142.
9. danah boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 43rd Hawaii International Conference on System Sciences* (HICSS '10), 1-10.
10. Christopher A. Cassa, Rumi Chunara, Kenneth Mandl, and John S. Brownstein. 2013. Twitter as a sentinel in emergency situations: lessons from the Boston marathon explosions. *PLoS Currents* 5, June.
11. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, 675-684.
12. 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*.
13. Thomas E. Drabek. 1970. Methodology of studying disasters: Past patterns and future possibilities. American Behavioral Scientist 13, 331-343.
14. Motahhare Eslami, Karrie Karahalios, Christian Sandvigt, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. "First I "like" it, then I hide it: Folk Theories of Social Feeds." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2371-2382. ACM, 2016.
15. Sasha Frere-Jones. 2012. Good things about Twitter. Retrieved May 26, 2016 from http://www.newyorker.com/culture/sasha-frere-jones/good-things-about-twitter
16. Huiji Gao, Geoffrey Barbier, Rebecca Goolsby, and Daniel Zeng. 2011. Harnessing the crowdsourcing power of social media for disaster relief. Arizona State Universtiy, Tempe, AZ.
17. Anthony Giddens. 1984. *The constitution of society: Outline of the theory of structuration*. University of California Press.
18. John Herrman. 2012. Twitter is a Truth Machine. Retrieved May 26, 2016 from http://gofwd.tumblr.com/post/34623466723/twitter-is-a-truth-machine
19. 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 the 11th International ISCRAM Conference*, 602-611.
20. Y. Linlin Huang, Kate Starbird, Mania Orand, Stephanie A. Stanek, and Heather T. Pedersen. 2015. Connected through crisis: emotional proximity and the spread of misinformation online. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), 969-980.

21. 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.
22. Willett Kempton. 1986. Two theories of home heat control. *Cognitive Science*, *10*(1), 75-90.
23. Pierre Lévy. 1997. *Collective Intelligence: Mankind's Emerging World in Cyberspace*. Plenum / HarperCollins.
24. Eden Litt. 2012. Knock, knock. Who's there? The imagined audience. *Journal of Broadcasting & Electronic Media* 56.3 (2012): 330-345.
25. Jim Maddock, Kate Starbird, and Robert M. Mason. 2015. Using historical twitter data for research: Ethical challenges of tweet deletions. *Workshop on Ethics at the 2015 Conference on Computer Supported Cooperative Work*, March, 14.
26. Alexis C. Madrigal. 2013. #BostsonBombing: The Anatomy of a Misinformation Disaster. Retrieved May 27, 2016 from http://www.theatlantic.com/technology/ archive/2013/04/-bostonbombing-the-anatomy-of-amisinformation-disaster/275155/
27. Alice E. Marwick & danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, no. 1 (2011): 114-133.
28. Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter Under Crisis: Can we trust what we RT?. In *Proceedings of the first workshop on social media analytics*, 71-79.
29. Onook Oh, Manish Agrawal, and H. Raghav Rao. 2013. Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly* 37, 2: 407-426.
30. 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.
31. Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1589-1599.
32. Enrico Louis Quarantelli. 1991. Radiation disasters: Similarities to and differences from other disasters. In *The Medical Basis for Radiation-Accident Preparedness III: The Psychological Perspective*, R. Ricks, M. Berger and F. O'hara Jr. (eds,). Elsevier Science Pub., New York, USA, 15-24.
33. Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World Wide Web*, 249-252.
34. Ralph L. Rosnow, James L. Esposito, and Leo Gibney. 1988. Factors influencing rumor spreading: Replication and extension. *Language & Communication* 8, 1: 29-42.
35. Ralph L. Rosnow. 1991. Inside rumor: A personal journey. *American Psychologist* 46, 5: 484-496.
36. Tamotsu Shibutani. 1969. Improvised news: A sociological study of rumor. *American Sociological Review* 34, 5: 781-782.
37. Emma S. Spiro, Sean Fitzhugh, Jeannette Sutton, Nicole Pierski, Matt Greczek, and Carter T. Butts. 2012. Rumoring during extreme events: A case study of Deepwater Horizon 2010. In *Proceedings of the 4th Annual ACM Web Science Conference*, 275-283.
38. 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 the iConference 2014*, 654-662.
39. Kate Starbird, and Leysia Palen. 2011. "Voluntweeters": self-organizing by digital volunteers in times of crisis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 1071-1080.
40. Kate Starbird, and Leysia Palen. 2010. Pass it on? Retweeting in mass emergenc*y*. In *International Community on Information Systems for Crisis Response and Management* (ISCRAM 2010).
41. James Surowiecki. 2005. *The Wisdom of Crowds*. Anchor.
42. Yuko Tanaka, Yasuaki Sakamoto, and Toshihiko Matsuka. 2012. Transmission of rumor and criticism in Twitter after the Great Japan Earthquake. In *Annual Meeting of the Cognitive Science Society*, 2387.
43. Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330, 6004: 686-688.
44. Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, 1395-1405.