Rumor Cascades

Adrien Friggeri
Facebook
friggeri@fb.com

Lada A. Adamic
Facebook
ladamic@fb.com

Dean Eckles
Facebook
deaneckles@fb.com

Justin Cheng
Stanford University
jcccf@cs.stanford.edu

Abstract

Online social networks provide a rich substrate for rumor propagation. Information received via friends tends to be trusted, and online social networks allow individuals to transmit information to many friends at once. By referencing known rumors from Snopes.com, a popular website documenting memes and urban legends, we track the propagation of thousands of rumors appearing on Facebook. From this sample we infer the rates at which rumors from different categories and of varying truth value are uploaded and reshared. We find that rumor cascades run deeper in the social network than reshare cascades in general. We then examine the effect of individual reshares receiving a comment containing a link to a Snopes article on the evolution of the cascade. We find that receiving such a comment increases the likelihood that a reshare of a rumor will be deleted. Furthermore, large cascades are able to accumulate hundreds of Snopes comments while continuing to propagate. Finally, using a dataset of rumors copied and pasted from one status update to another, we show that rumors change over time and that different variants tend to dominate different bursts in popularity.

Introduction

Social ties have always played an important role in disseminating relevant information through a social network, whether it be factual, false, or questionable. In social media, the ease of information dissemination through weak ties (Bakshy et al. 2012; Granovetter 1973), people's trust of information from their friends, and short path lengths (Ugander et al. 2011) together create an environment where information can spread quickly throughout the world. In the absence of reliable official or established sources, individuals can use these networks to share important information quickly and coordinate action (Starbird and Palen 2012). In such situations, rumors may be the best information available. More generally, sharing rumors and gossip can also build and maintain social ties, as individuals provide early information to peers, express common perspectives and affiliations, and cope with uncertain circumstances (Allport and Postman 1947). Online communication technologies are often neutral with respect to the veracity of the information: they can facilitate the spread of both true and false information. Nonetheless, the actions of the individuals in the network (e.g., propagating a rumor, referring to outside sources, criticizing or retracting claims) can determine how rumors with different truth values, verifiability, and topic spread.

Online, where many rumors can spread over the same network, observing a large corpus of such rumors allows studying both common patterns and heterogeneity in their propagation. This includes variation in propagation of rumors of differing veracity: Do true or false rumors spread faster and further? Or are rumors with complex, disputed, or unknown truth values most contagious? Since the same communication technologies are used to spread non-rumors, rumor cascades can be compared with other types of cascades over a common network and affordances for propagation.

Online media are often not only the site of propagation of information, but also of persistent discussions about that information. Individuals encountering a rumor may turn to other sources to understand, evaluate, debunk, or bolster a rumor, often depending on their prior beliefs (Lewandowsky et al. 2012). Online media afford both searching for this information and referring to it alongside the rumor itself. In particular, some information of dubious veracity shared via social media is so contagious that there are dedicated sites, such as *Snopes.com*, that document the spread of the rumor, as well as try to determine its truth value. We expect that references to outside resources, such as Snopes, play a substantial role in variation in how quickly a rumor spreads; they also support the construction of large corpora of rumors.

In this paper, we examine the spread of rumors on Facebook. To this end, we consider two different technological affordances for rumor propagation on Facebook: uploading and resharing of photos and copying-and-pasting of text as a text post. Each allows us to construct the near exact path that the rumors take across the social network – the diffusion of rumors being something that is ordinarily not easy to trace. We measure the replication and longevity of instances of each rumor within Facebook, as well as the role of references to outside sources to this success. We complete the analysis by studying the modification of rumors using textual data.

Related work

Despite the importance of the propagation of rumors and the opportunities for their study presented by online social networks, little is known about rumor propagation on these networks. While information diffusion in online social networks has recently been the subject of considerable scholarly attention, this work has generally not made use of a distinction between true and false information. Whether studying the spread of news (Bhattacharya and Ram 2012), characterizing the structure of information cascades (Dow, Adamic, and Friggeri 2013; Goel, Watts, and Goldstein 2012; Liben-Nowell and Kleinberg 2008), or estimating influence on the diffusion process (Bakshy et al. 2011), the veracity of information is generally not included in these quantitative analyses.

Some recent work has begun to study rumors and hoaxes in social media.[1] Similarly to previous psychological research (Prasad 1935), one strategy has been to examine a specific event that is the subject of rumors: Kaigo (2012) traced the popularity of a rumor and counter-rumor regarding the Cosmo Oil refinery fire following the 2011 Great East Japan earthquake, while Oh et al. (2010) produced evidence for the role of anxiety and informational ambiguity in the spread of rumors about the 2010 Haiti earthquake. Mocanu et al. (2014) studied interactions around reliable and less reliable sources of Italian political news on Facebook. We instead examine a large corpus of rumors, both true and false, concerning many independent topics. Other work has focused on identifying rumors in networks: Kwon et al. (2013) identified temporal, structural, and linguistic features of rumors on Twitter, Gupta et al. (2013) tried to predict whether images being transmitted on Twitter were real or fake; while Qazvinian et al. (2011) attempted to predict whether tweets were factual or not, while also identifying sources of misinformation. We instead rely on structured external resources to classify rumors and incorporate this information in our analysis of propagation dynamics. We additionally analyze how rumors can mutate, with these mutations influencing and competing with each other.

Recent work within political psychology has built on theories of motivated reasoning to examine how misinformation spreads and how to combat it (Lewandowsky et al. 2012). Even in traditional news environments, misinformation can become widespread belief that is resistant to correction, whether because of limited distribution of corrections, selective exposure to partisan media (Iyengar and Hahn 2009), or motivated reasoning when processing the correct information (Nyhan and Reifler 2010; Nyhan 2010).

Collecting rumors

To track the spread of rumors on Facebook, we need two types of information: a corpus of known rumors, and a sample of reshare cascades circulating on Facebook which can be matched to the corpus. The website *Snopes.com* has documented thousands of rumors, and provides the starting point for our data collection.

Rumors documented by Snopes

We retrieved from the Snopes website two classifications of the rumors they have covered and analyzed. The first is the veracity, which includes "true" and "false", but also a range of intermediate or orthogonal values, i.e. partly true, multiple truth values, unclassifiable, undetermined, and legend. We also retrieved the broad thematic category Snopes assigned to the rumor, e.g. Politics, Food, "Fauxtos", etc. After sanitizing the corpus — merging duplicate entries and removing entries with contradictory information — there remained 4,761 distinct rumors.

Around 22% of those rumors are related to politics, 12% are either "photoshopped" images or real photos with fake backstories (*Fauxtos*) and 11% of rumors fall into a broad category called *Inboxer Rebellion* which consists of stories of emails and chain letters of "dubious origin and even more dubious veracity". There is then a long tail of rumors ranging from 9/11 to rumors specifically about Coca-Cola (Figure 1).

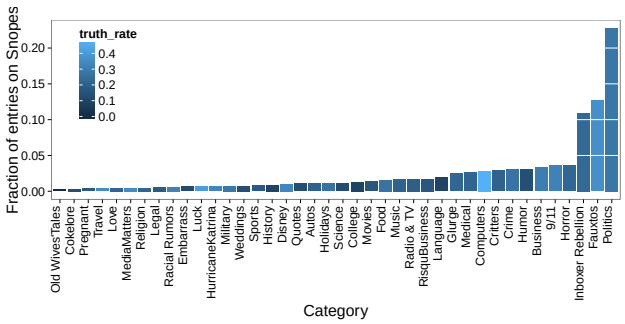

Figure 1: Distribution of number of rumors among categories on Snopes, along with the fraction of these stories that are evaluated as true.

For simplicity's sake, we recode the veracity in one of three possible values: 45% of the Snopes corpus consists of rumors that have been debunked and are thus considered false, 26% turn out to actually be true, and we place the remaining 29% (*multiple truth values, mixed, undetermined*) into a group of rumors that are "maybe" true — either because parts of the rumor are true whereas others aren't, or because the Snopes team was unable to verify the story. This points to a tendency of Snopes to document more false stories than true ones, in line with the stated mission on the front page of their website: "*Welcome to Snopes.com, the definitive Internet reference source for urban legends, folklore, myths, rumors, and misinformation*".

How information propagates on Facebook

Rumors readily propagate through whatever medium is available to them: word-of-mouth, email, and before email, even xerox copies (Bennett, Li, and Ma 2003). With each technological advancement that facilitates human communication, rumors quickly follow. One such change was the introduction of the 'Share' button, which accompanies any

[1] In some literatures, the term 'rumor' has also been used as a generic term to describe any type of information propagation in a network (e.g., (Kostka, Oswald, and Wattenhofer 2008)).

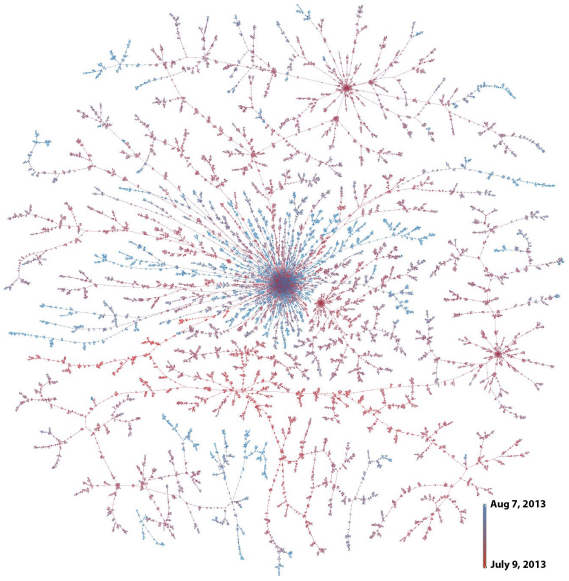

Figure 2: Cascade of reshares of a Cabela's sporting goods store receipt attributing addition sales tax to "Obamacare". Shares which did not prompt further shares are excluded. The coloring is from early (red) to late (blue).

status update, link, or photo posted on Facebook. It allows viewers of the content (e.g., friends and followers of the creator) to share the post. If the content was originally posted publicly, meaning that it can be viewed by anyone, the share exposes the sharer's friends and followers to the post as well, who in turn can again share the content.

For information that is sufficiently viral, in the sense that many individuals are likely to share it, large information cascades can occur — these consisting of the original post and the tree of reshares. Since, on Facebook, the audience of the content can never be expanded beyond what is specified by the root node, we only consider cascades of publicly viewable content. In the first part of the analysis, we restrict our attention to content in the form of photos, which comprise the majority of share cascades on Facebook (Dow, Adamic, and Friggeri 2013). The rumor can be contained in the photo itself, or can be added as the photo's caption, or spans both.

We supplement the photo cascades dataset with a collection of rumors propagating on Facebook prior to the introduction of the 'Share' button, through a copy-and-paste replication mechanism. Textual memes can be modified in addition to replicated, and we examine both the evolution of the rumor, and the counter-rumor for two specific examples.

Sampling rumors propagating as photos

To identify rumor cascades propagating as photos, we rely on user-generated information posted in the form of a Snopes link embedded in a comment associated with either the original photo or one of its reshares. Such comments are posted by individuals to either warn their friends that something they posted is inaccurate or to the contrary, to validate that a rumor, though hard to believe, is in fact true.

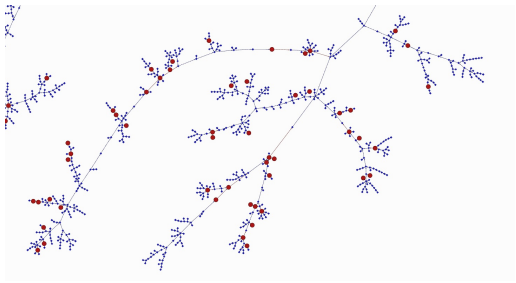

Figure 3: A branch of the cascade shown in Figure 2. Shares that received a Snopes link are displayed in red.

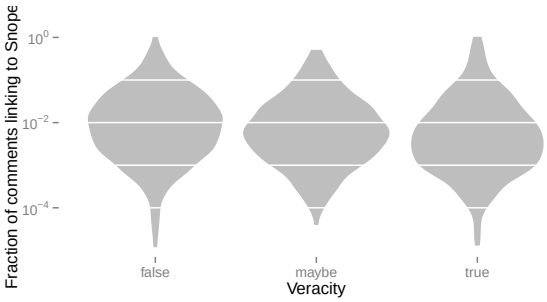

Figure 4: The fraction of comments linking to Snopes depends on the veracity of the rumor. Commenters are more likely to point out that a rumor is false than true.

We gathered a sample of 249,035 comments on either photos or shares of photos, posted during July and August 2013 and containing a valid link to a rumor covered by Snopes. From those, we then tagged 16,672 individual cascades, containing 62,497,651 shares. It is possible that a single photo cascade receives comments with several different Snopes links, either because the commenters identified the rumor incorrectly, or because the photo propagating is a mixture of rumors. To limit our dataset to just photo cascades representing a single rumor, the above sample contains only photo cascades with a single dominant rumor where more than 95% of the links point to the same Snopes article. This excludes 1.9% of the cascades.

It is important to note that this sample is heavily biased since it relies on either the original photos or one of its shares receiving a comment linking to Snopes. Given that the probability p_s of receiving such a comment on an individual photo or share is very small — between 0.1% and 0.3% of comments contain a link to Snopes — this means that we are capturing only a small fraction of all rumor uploads, and furthermore, we are more likely to find larger cascades in our sample, since each share increases the probability that one of these comments is received. To complicate matters further, the veracity and category of the rumor can play a role in whether a share receives a Snopes comment, as for example false rumors elicit more Snopes links than true ones (Figure 4). This presents challenges in inferring aggregate characteristics, primarily the popularity, of the entire population of uploaded rumors.

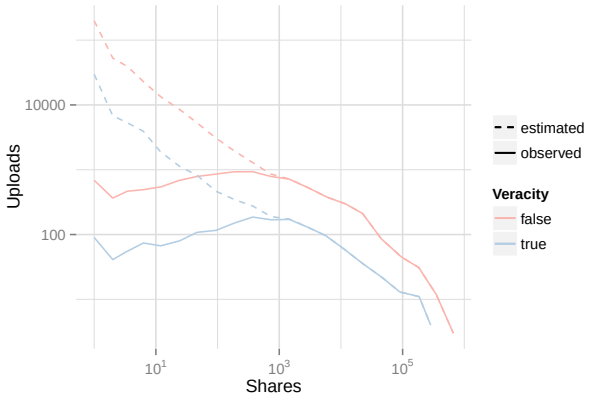

Figure 5: Distribution of the number of shares of uploads before and after estimation, for both true and false rumors.

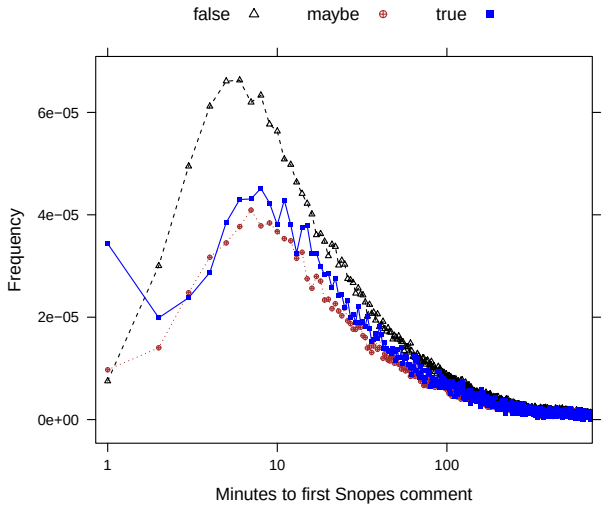

Figure 6: Distribution of time from when a reshare is posted to when it is snoped. False rumors are snoped more frequently, most notably shortly after the reshare is created.

The probability that a cascade consisting of n photos and shares is detected is approximately $1-(1-p_s)^n$ where p_s is the probability of receiving a comment containing a link to Snopes. That probability varies between stories, due to a combination of factors: whether the rumor is true or false, whether it is controversial enough for readers to want to know more and the ease with which one can end up reading the Snopes page after searching for keywords used in the story. For example consider a false rumor claiming that the best way to distinguish two-way mirrors from ordinary mirrors is looking for the absence of a gap between an object and its reflection,[2] that story was shared 140,331 times and overall received 20,948 comments, 1,346 of which referenced a Snopes article.

In contrast, a photograph of an old 'money bags' text meme claiming "*This year July had 5 Fridays, 5 Saturdays, and 5 Sundays. This happens once every 823 years. This is called money bags...*", was shared 1,259,642 times and received 174,728 comments, only 908 of which linked to Snopes — perhaps that photo was too obviously an old email chain letter to prompt individuals to search for whether it was true or false. Incidentally, the claim is false in 2013, but was true in 2011, at least the part about July having 5 Fridays, Saturdays and Sundays. The 823 year claim is false, while the money bags aspect, well, has yet to be verified.

However, while individual stories do vary in their likelihood of receiving a comment linking to Snopes, one can still do a rough estimate of the total number of rumors that are uploaded from the ones that do get detected that way – essentially, we infer the head of the distribution from its tail. To estimate p, the rate of receiving snopes, we examine the proportion of shares receiving a Snopes comment per cascade of size larger than 10^4 – even with an extremely low estimate of 0.05% of comments mentioning Snopes, the probability of such a cascade being undetected would be less than 0.007, so we can assume that the fraction of comments linking to Snopes on these large cascades is a good estimate of p_s. We do so separately for True, False and Maybe (True) stories, given that, as we have seen earlier, these tend to elicit different responses from commenters (Figure 4). From this we obtain $p_{\text{true}} = 3.03 \times 10^{-3}$, $p_{\text{false}} = 3.46 \times 10^{-3}$, $p_{\text{maybe}} = 3.68 \times 10^{-3}$ and we can reconstruct an estimate of the real distribution of the number of shares (Figure 5), showing the more familiar pattern of many rumor instances not gaining much traction but a few gaining a lot.

Not only are false rumors more likely to receive snopes comments, they are also more likely to do so shortly after being posted (Figure 6). Casual inspection of the cascade depicted in Figure 3 might suggest that being snoped results in a reshare having fewer children. In the following sections we will quantitatively examine the association of a reshare being snoped with such outcomes.

Structure and dynamics of rumor cascades

We set to explore several aspects of rumor dynamics, from the initial entry of a rumor on the Facebook ecosystem via a photo upload, to the cascade that is generated from that original upload, to the reactions the rumor evokes in those who are exposed to it. Furthermore, we wish to examine the effect that reactions including Snopes links have on the propagation of the rumor, both in terms of whether a reshare of a rumor is removed, and whether its ability to induce further reshares is affected.

Presence and virality by category

Compared to the number of photos uploaded on Facebook some categories are either over- or under-represented on Snopes (Figure 7). For example, Political stories amount to 32% of cascades whereas only 22% of Snopes stories are in that category; similarly, Food, Crime and Medical photos are uploaded more than we would expect. Conversely, the fraction of photos about 9/11 is a lot lower than the fraction

[2] http://www.snopes.com/crime/warnings/mirror.asp

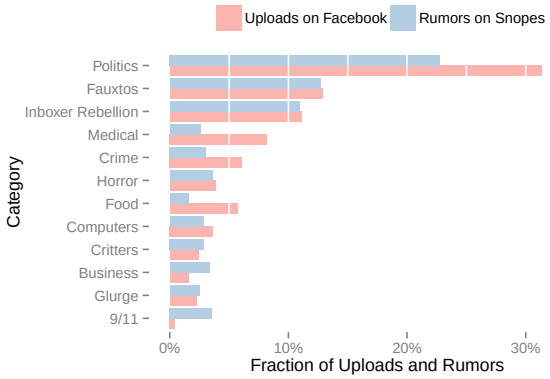

Figure 7: Some categories, such as Politics, Medical or Food are over-represented on Facebook compared to what would be expected if they were randomly drawn from the Snopes corpus.

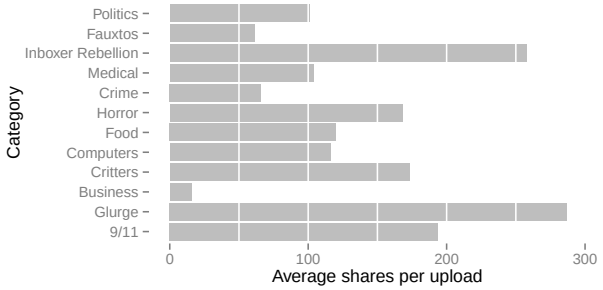

Figure 8: Average number of shares per upload for several categories. Comparing these values to those in Figure 7, it is striking that virality and number of distinct cascades are unrelated.

of rumors in that category, which can be explained by its historical nature: those rumors circulated in 2001 and were subsequently covered by Snopes but have since stopped propagating at such a large scale.

As we mentioned in the dataset description, around 45% of rumors covered on Snopes are false, which has to be contrasted with the 62% of cascades on Facebook that are tagged as false. Similarly, only 9% of cascades are true on Facebook, whereas 26% of the Snopes corpus was found to be true.

Although false rumors are predominant, we observe that true rumors are more viral — in the sense that they result in larger cascades — achieving on average 163 shares per upload whereas false rumors only have an average of 108 shares per upload. When looking at the categorical level, those that are the most popular in terms of number of distinct cascades do not appear to be the more viral: consider for example the Fauxtos category which mostly contains fake photos, despite being the second most popular category has cascades of size on average 61, whereas rumors in the much less popular categories Glurge (motivational images)

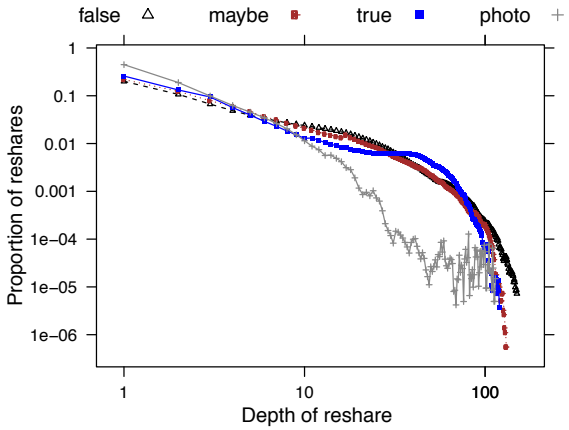

Figure 9: Proportion of reshares at a given depth in the larger photo resharing cascade. Reshares from rumor cascades are compared with a reference sample of reshares from other photo cascades, matched by vigintiles of cascade size for cascades with 100 or more reshares. Rumor cascades tend to be deeper, in that more reshares are at greater depths, than the reference cascades.

and Inboxer Rebellion (chain letters) are shared more than 250 times per upload.

Cascade characteristics

Photos can be uploaded by two types of entities on Facebook: users and pages. Users have reciprocal friendship ties with other users, and they can optionally allow other users to follow them by subscribing to their public updates. Users can select the audience for each post, whether it be just a subset of their friends, all of their friends, or anyone. Pages on the other hand can be managed by several users, have just followers, and all their posts are public.

Because the audience of a photo can never be expanded, a photo shared with friends, even if "reshared" by a friend, will only still be visible to the friends of the original poster. Therefore our dataset, which skews toward cascades large enough to have been snoped, consists of publicly shared photos. In addition to being posted as public photos, most large cascades on Facebook are also uploaded by pages or users with many friends and/or followers. For example, among all photo cascades of 100 or more reshares in July and August, 94.2% were initiated by pages. In contrast, pages play a lesser role in propagating the Snopes cascades in our sample. They were responsible for half of false rumors (49.3%) and maybe true (49.9%), but initiated a somewhat higher 62.6% of cascades of true rumors.

One might suspect then that rumors are more likely to be inserted into the network by less well-connected entities, either users or pages. In order to reach comparable size to a typical large cascade, the rumor has to go deep. Measuring depth requires knowledge of the path the information took, which we constructed by rechaining clicks, impressions, and connection edges as described by Dow, Adamic, and Friggeri (2013). We do the same for all photos with 100 or more

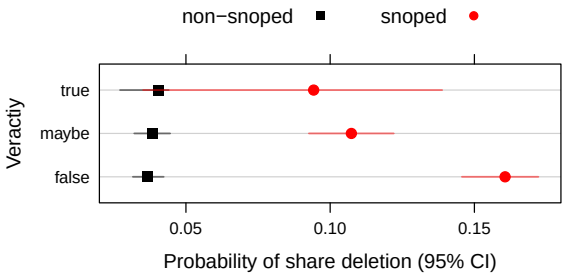

Figure 10: Probability that a reshare will be deleted as a function of the veracity of the rumor and whether it was snoped.

reshares uploaded during the prior month.

The resulting distribution of cascade depth shows rumors to be more viral than photo cascades in general, consistent with the propagation relying less on who posts the rumor, and more on the highly contagious nature of the rumor itself.

Reshare deletion

Individuals propagating rumors may attempt to disrupt their role in propagation or disassociate themselves from the rumor if they learn that it is false or otherwise experience social conflict as the result of propagating it. One opportunity for the resharer to reconsider is when others comment on the reshare by referring to outside sources where the veracity of the rumor is discussed, e.g. when they snope the reshare. In the case of false rumors, we hypothesized this would sometimes cause individuals to delete the reshare, perhaps because of a desire to not propagate or be associated with false rumors. This could result in a higher observed deletion rate for snoped reshares.[3]

This association could also be produced by many other confounding processes, such as differences in deletion rates across rumors associated with their frequency of being snoped. While it is difficult to adjust for all such factors, we stratify our analysis on the Snopes URL and then combine these results, weighting by the number of snoped reshares. If conditioning on URL makes snoping ignorable, then the difference between deletion proportion are unbiased estimates of the average treatment effect on the treated (i.e., the average effect of snoping on shares that get snoped).

Consistent with our hypothesis, reshares about false rumors are 4.4 times (95% CI [3.7, 5.2]) as likely to be deleted when snoped than when not. In fact, for all three veracity categories, we estimate snoped reshares are more likely to be deleted (Figure 10). This difference was statistically significant for maybe (true) rumors ($p < 0.001$) but not quite so for true rumors ($p = 0.079$).

[3]This association could also be produced by many other processes, such as resharers with more active friends being more likely to delete their own content. We observe deletions for 28 days from the reshare. All statistical inference in this section uses confidence intervals and p-values from an online half-sampling bootstrap (Owen and Eckles 2012) clustered on URL.

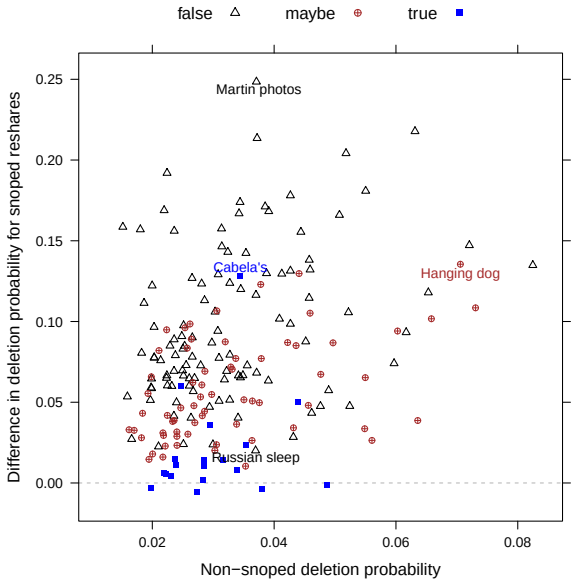

Figure 11: Estimated probability difference for snoping on deletion for each URL as a function of deletion probability for non-snoped reshares. Rumors with the largest estimated difference in each veracity status are labeled, as is the false rumor with the smallest difference.

Note that Snopes entries for maybe rumors and even true rumors can contain information that might contradict a particular instance of the rumor; likewise, the Snopes entries may highlight the age of the rumors, making the poster aware it is older than they previously believed. Despite higher deletion rates for snoped reshares across all veracities, the difference between deletion probabilities for snoped and non-snoped reshares is larger for false rumors than either true ($p = 0.026$) or maybe ($p < 0.001$) rumors. Even though false rumors have a higher deletion rate even when not snoped, they also have a higher relative risk of deletion when snoped than true rumors ($p = 0.03$) and maybe rumors ($p = 0.0011$). Within snoped reshares, reshares about false rumors are more likely to be deleted than those about either true ($p = 0.032$) or maybe ($p < 0.001$) rumors.

To further examine heterogeneity in deletion rates and potential effects of snoping for different rumors, we fit a logistic regression model predicting deletion and incorporating random effects for each URL. Figure 11 displays the estimates for URLs with at least 100 snoped and non-snoped reshares. There is substantial variation in deletion probabilities, even within rumors with the same ostensible veracity.

We consider some extreme examples that illustrate plausible causes of this heterogeneity. The two false rumors with the most extreme snoping effects are one that involves claims about photos of Trayvon Martin and his physical characteristics[4] and one about a experiment on sleep that supposedly took place in Russia in the 1940s.[5] The former

[4]http://www.snopes.com/photos/politics/martin.asp

[5]http://www.snopes.com/horrors/ghosts/russiansleep.asp

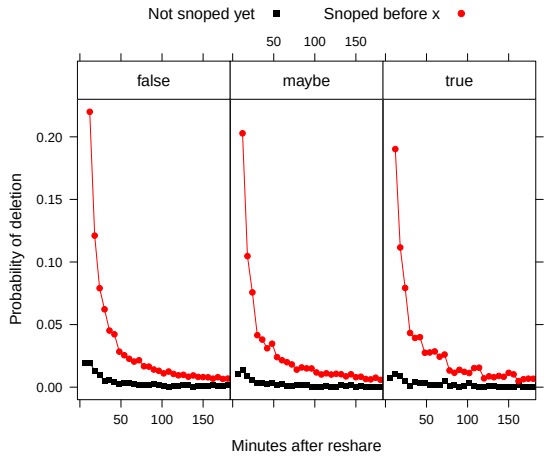

Figure 12: Probability that a reshare will be deleted some time as a function of whether it has been snoped prior to that time. Each point is a 6-minute interval. Reshares that are snoped shortly after posting are very likely to be deleted.

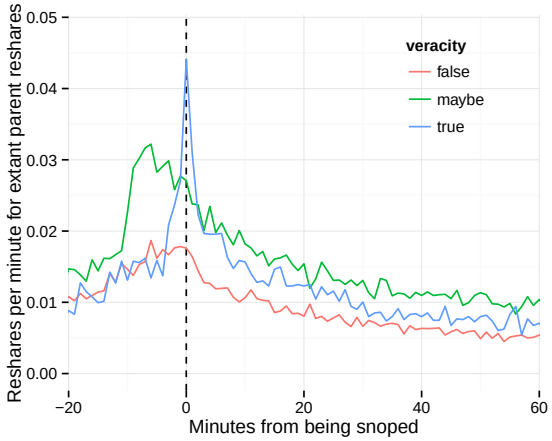

Figure 13: Reproductive rate of reshares that are eventually snoped by time from being snoped. The act of snoping coincides with a boost for true stories, but appears to be detrimental in the short run to the propagation of the false cascade.

concerns a criminal trial that attracted substantial contemporaneous attention. Some versions of this rumor involved mistaking a photo of a rapper for a photo of Martin (perhaps an embarrassing error). Additionally, snoping may indicate that the resharer's friends may have disagreed with the resharer about broader issues in the trial; that is, this estimate may reflect confounding of snoping effects with heterogeneous reshare neighborhoods.

Among maybe rumors, one about a photo of two young men hanging a small dog has the highest snoping effect on deletion; standard versions of this post urge viewers to help try to identify the young men.[6] Snopes reports these photos are from Malaysia, so it could be that resharers decided that they and their friends are not in a position to identify the men. These reshares also have a high baseline rate of deletion, perhaps reflecting resharers' desires to remove a displeasure-inducing photo from their profiles.

Finally, the rumor classified as true with the highest estimated snoping effect is the Cabela's medical device tax rumor previously depicted in Figures 2 and 3. While this rumor has multiple claims identified as true by Snopes, one of its more consequential conclusions is identified as false.[7] Thus, this case may primarily reflect the difficulties in automatic categorization of the veracity of rumors that involve multiple independent claims.

By comparing reshares that did or did not get snoped, the preceding analysis provides some evidence that references to external sources refuting a rumor cause the resharer to delete their reshare. For many cases, a reshare being snoped is associated with high deletion probabilities; thus, if sufficiently common, such references could play a substantial role in differentially slowing the spread of false, compared with true, rumors. Further evidence for this effect is provided by considering only reshares that are snoped and noting that reshares snoped by some time after resharing are more likely to be deleted at that time (Figure 12).

Effects of snoping on resharing

While causing a reshare to be deleted can stem the spread of a rumor, comments containing Snopes links also have the potential to retard the ability of the reshare to replicate, even if the reshare itself is not deleted. A user seeing the reshare in their News Feed might also notice it was snoped, which could affect their choice to reshare it. To examine how snoping affects the replicative success of a reshare, we consider reshares that received a comment containing a Snopes link (i.e., were snoped) within the first day and analyze the frequency with which *child reshares* occur from this reshare over time. A child reshare is one that was produced by clicking 'Share' on the parent reshare directly, or one that was likely created by a user being exposed to the parent reshare. We then compare child reshare rates over time, relative to the time of snoping.[8]

In this dataset of snoped reshares, most child reshares and Snopes comments are proximate in time, as the period of time over which a post receives peak attention is relatively short. The median separation between a parent reshare and child is 1.9 hours, while the mean is 5.8. The time to the first snoping is a bit shorter, a median of 1.3 hours and a mean of 3.8. We analyze the data binned by minutes between the first Snopes comment and each child reshare. A negative value occurs when a child reshare occurs before the Snopes com-

[6] http://www.snopes.com/critters/crusader/hangingdog.asp

[7] http://www.snopes.com/politics/taxes/medicaldevice.asp

[8] An alternative analysis would include reshares that were not snoped, but we expect these reshares differ from snoped reshares in numerous ways; this is one motivation for the interrupted time series analysis presented here. We also excluded reshare cascades where the root had been snoped, to measure just the effect of the reshare being snoped directly.

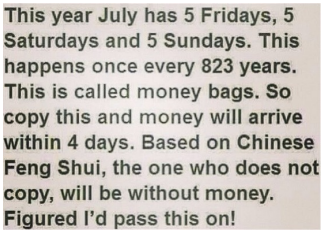

Figure 14: The 'money bags' meme in photo form.

ment. For each bin we compute the number of reshares that occurred in that bin per parent reshare that existed as of that time. This reproductive rate is plotted in Figure 13, which shows interesting differences between reshares by veracity. For reshares of false or dubious truth value, the reshares peak during the 10 minutes before the reshare is snoped. Then once the reshare is snoped, the likelihood of resharing falls. On the other hand, for reshares of true rumors, the peak coincides with being snoped. However, in the longer term, there appears to be little effect. In fact, while 45.2% of reshares of true stories occur after being snoped for the first time, potentially also due to the lower rate of snoping true comments, an even higher proportion are reshares occur after the first Snopes comment on a false or mixed reshare, 51.9 and 59.4 percent respectively. This suggests that their high virality can overcome a temporary setback dealt by being snoped.

That snoping has little long-term effect is consistent with not all comments being read by users before resharing, either due to lack of interest, or because other, more recent comments are more easily viewable.

Rumor evolution

Although our analysis has so far focused on share cascades of photos, some of those are little more than screenshots or scans of textual memes that have been circulating through various forms of communication for years.

Variants and burstiness

For example, a meme called "money bags", promising money to those who propagate it, shown in Figure 14, had a copy posted on July 15th, which by September 24th was shared 1,125,055 times. The lucky auspices advertised in the meme is that 'This July' has 5 Saturdays, 5 Sundays, and 5 Mondays, and that this happens once every 823 years. First, this is not true for 2013, the year in which it is posted and being shared, but it was true in 2011. Second, even for variants of the meme that are true (at least about the calendar event, if not the arrival of money), the frequency of occurrence is not every 823 years. October 2010 had 5 Fridays, Saturdays, and Sundays (5 FSS), while in 2011 it was true in July, and in general on average one month each year has 5 FSS. Remarkably, even as the month and other details of the meme change, the "823 years" false part of the meme remains intact.

In order to trace the change of a meme, we use an older dataset of copy and paste memes, which encompass all memes posted on Facebook as anonymized status updates between April 2009 and mid-October 2011. This period predates the 'Share' button, and memes spread within Facebook primarily by being copied and pasted from one status update to another. Unsurprisingly, memes which occurred in significant numbers nearly always contained some form of replication instructions, e.g. "copy and paste", "repost", or "make this your status". We used these phrases as a filter to generate a dataset of potential memes, tokenized them into 4-grams – 3 grams created more noisy clusters – and applied an agglomerative clustering technique based on cosine similarity. This yielded over $6,000$ meme clusters during the period, one of them corresponding to the "money bags" meme.

Figure 15 shows that the popularity of the meme is highly bursty, with significant lulls during which almost no copies of the meme were posted. Crucially, the meme never quite dies out. Rather it persists in low frequencies even months after it was posted, creating the potential for flare-ups once the conditions are right. For example, although a few status updates noted in 2010 that the next 5FSS occurrence will be in July of 2011, what seems to fuel the July 5FSS variant is actually a flare up of the October 5FSS variant which occurred in January 2011, at which point it was no longer correct. Another correct variant, October 5SSM appears at this time, but has much more limited spread, until October, when it flares up, seemingly in response to a peak of the October 5FSS version. During this period, there were hundreds of status updates debunking the rumor in various ways, e.g. pointing out that the next occurrence of 5FSS after October 2010 would be July 2011, then followed by March 2013, etc., along with skepticism even from those who were passing the original rumor on, e.g. "Don't know whether to believe this, but here goes…". The status updates debunking the rumor were however not themselves viral, they lacked incentive – they did not promise bags of money – and often did not ask to be propagated.

Counter-rumors

There are examples of counter-rumors spreading in the heels of a rumor: in July of 2011, a rumor started circulating that Facebook would start charging for access: "It's official [. . .] It even passed on TV. Facebook will start charging this summer. If you copy this on your wall your icon will turn blue and Facebook will be free for you. Please pass this message. If not your count will be deleted. P.S. This is serious. The icon turns blue so please put this as your status". This appears to have been the first significant English variant of the rumor (a Danish variant protesting a fictitious plan to convert Facebook to a for-fee site had been copied thousands of times in April of 2009). A second English-language variant of the original rumor, purporting a price grid was about to be instituted appeared two months later, with very little activity in-between.

Figure 16 shows different variants of both the rumor and counter rumors. Counter rumors in the form of informational messages debunking the hoax such as "don't blindly copy and paste warnings just because your Facebook friend's status tells you to do so. Although you probably mean well, you could be helping a hoax become more popular", achieved only modest popularity, about 5,000 copies in total. In con-

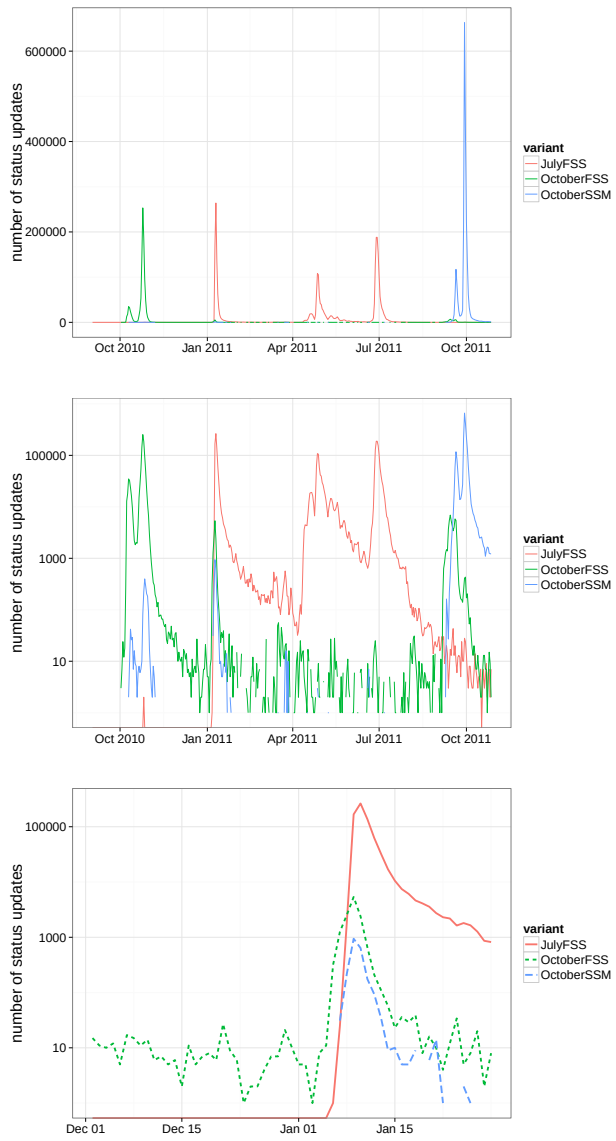

Figure 15: (a) Popularity of the *money bags* meme over time, with the following variants: October 5FSS (correct in 2010), and July 5FSS and October 5SSM , both correct in 2011. (b) the same but on a log-scale to accentuate the slow decay of the peaks (c) popularity trends in January 2011.

trast, a humorous parody was widely copied: "On September 31st 2011 Facebook will start charging you for your account. To avoid this you must get naked, stand on your dining room table and do the Macarena all the while singing 'I will survive' after filming and posting it to your Facebook wall and YouTube then and only then will Mark Zuckerberg come down your chimney to tell you that your account will stay free. Pass it on it must be true because someone on Facebook I hardly know told me." Interestingly, this parody was present during the first rumor peak, but didn't dominate the rumor until the second burst in rumor activity. During

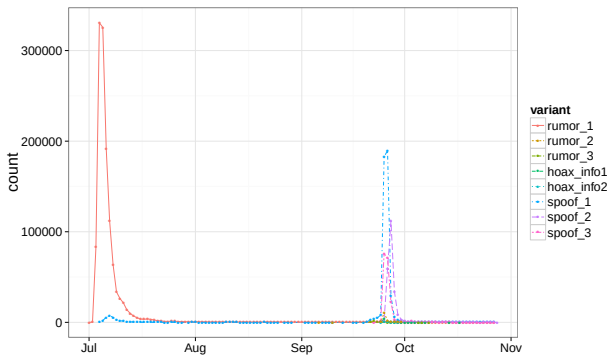

Figure 16: Popularity over time of the "Facebook will start charging" meme and its variants.

the second peak, two additional parodies also achieved wide distribution. The end result was that the three most popular spoofs achieved over 8.6 million posts, close to the 12.7 million the original rumor garnered. A majority (65.2%) of those resharing the most popular spoof had at least one friend who had posted the original, thus they were probably likely to appreciate the humor.

The above examples demonstrate the burstiness, recurrence, and even mutability of rumors on Facebook. Sometimes the changes are innocuous, e.g. updating the month in the money bags meme at the new year, but other times they are intentional. And it is the intentional changes that speak to the endless creativity that can be garnered in the social network, e.g. one spoof, copied over 300 times starts with "Today I have decided to start charging for the rare and wonderful opportunity of being my Facebook friend [...] Copy and paste this message along with your credit card information [...]". In the case of the moneybags meme, the network voted, by way of propagation, for the correct variant in each year. It is unclear how the shift to resharing photos rather than copying and pasting text will affect the evolution of rumors. It could be a coincidence that the most widely reshared rumor in July 2013 was an incorrect variant, or it may be a consequence of the photo not being easily modifiable to become the correct March 5FSS variant instead.

Conclusion

In this study we showed how readily and rapidly information can be transmitted via social ties, even when it is of dubious veracity. We traced a population of rumors being inserted via photo uploads into the Facebook social network, which were revealed as such when someone posted in response a comment with a link to Snopes. We examined the cascades resulting from individuals resharing the photo, finding that some rumors proceeded to thrive in the environment, broadcasting from one friend to the next. These cascades ran deep, deeper than the average photo reshare cascade of similar size, and elicited different responses depending on their veracity.

False rumors were frequently uploaded, and also frequently snoped. However, it was the true rumors that were

most viral and elicited the largest cascades. We note that although reshares of a rumor that have been snoped are several times more likely to be deleted, the cascade overall easily continues to propagate, as there are many more non-snoped reshares than snoped ones. We further find that for false and mixed rumors, a majority of reshares occurs after the first snopes comment has already been added. This points to individuals likely not noticing that the rumor was snoped, or intentionally ignoring that fact.

We further find the popularity of rumors — even ones that have been circulating for years in various media such as email and online social networks — tends to be highly bursty. A rumor will lie dormant for weeks or months, and then either spontaneously or through an external jolt will become popular again. Sometimes the rumors die down on their own, but in one particular case of a rumor claiming that Facebook would start charging a fee, we observed the rumor being dwarfed by an antidote more powerful than the truth: humor.

This study has important limitations. Most of our analysis focused on rumors propagating on Facebook through photos and their captions, which were identified as rumors by comments linking to Snopes. The specifics of resharing mechanisms on Facebook contributes to the observed data in unknown ways. Likewise, this method for collecting rumors produces a biased sample; we examined some aspects of this bias, but cannot credibly fully correct for it. Both of these limitations should motivate the analysis of rumors spreading through different mechanisms and collected through different means, as we did in the final section of the paper. Our analyses of the effects of receiving comments linking to Snopes are purely observational, and so can suffer from confounding bias. In the analysis of deletion, we adjusted for differences by Snopes URL, while in the analysis of child reshares, we used an interrupted time series, but important biases may remain.

The bursty nature of rumors remains a mystery. It would be interesting to examine whether rumor flare-ups are fueled by the presence of individuals who have never been exposed to the rumor, or whether, to the contrary, the rumor relies on those who know it well to retell it when prompted. In this paper we only scratched the surface of the question of whether there are subpopulations in the social network particularly susceptible to rumors, and others who are more likely to snope. Furthermore, it is unclear how individuals change their attitudes toward rumors, and whether being snoped or reading a comment containing a Snopes link would make an individual more or less likely to propagate subsequent rumors. These and other questions we leave for future work.

References

Allport, G. W., and Postman, L. 1947. *The Psychology of Rumor.* Henry Holt.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: Quantifying influence on Twitter. In *Proc. WSDM'11*.

Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The role of social networks in information diffusion. In *Proc. WWW'12*, 519–528.

Bennett, C. H.; Li, M.; and Ma, B. 2003. Chain letters and evolutionary histories. *Sci. Am.* 288(6):76–81.

Bhattacharya, D., and Ram, S. 2012. Sharing news articles using 140 characters: A diffusion analysis on Twitter. In *Proc. ASONAM'12*, 966 –971.

Dow, P. A.; Adamic, L. A.; and Friggeri, A. 2013. The anatomy of large Facebook cascades. In *Proc. ICWSM'13*.

Goel, S.; Watts, D.; and Goldstein, D. 2012. The structure of online diffusion networks. In *Proc. EC'12*, 623–638.

Granovetter, M. 1973. The strength of weak ties. *AJS* 78(6):1360–1380.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. In *Proc. WWW'13*, 729–736.

Iyengar, S., and Hahn, K. S. 2009. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication* 59(1):19–39.

Kaigo, M. 2012. Social media usage during disasters and social capital: Twitter and the Great East Japan earthquake. *Keio Communication Review* 34:19–35.

Kostka, J.; Oswald, Y.; and Wattenhofer, R. 2008. Word of mouth: Rumor dissemination in social networks. In *Structural Information and Communication Complexity*, volume 5058. 185–196.

Kwon, S.; Cha, M.; Jung, K.; Chen, W.; and Wang, Y. 2013. Aspects of rumor spreading on a microblog network. In *Social Informatics*, volume 8238. 299–308.

Lewandowsky, S.; Ecker, U. K.; Seifert, C. M.; Schwarz, N.; and Cook, J. 2012. Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest* 13(3):106–131.

Liben-Nowell, D., and Kleinberg, J. 2008. Tracing information flow on a global scale using Internet chain-letter data. *PNAS* 105(12):4633.

Mocanu, D.; Rossi, L.; Zhang, Q.; Karsai, M.; and Quattrociocchi, W. 2014. Collective attention in the age of (mis) information. *arXiv preprint arXiv:1403.3344*.

Nyhan, B., and Reifler, J. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32(2):303–330.

Nyhan, B. 2010. Why the 'death panel' myth wouldn't die: Misinformation in the health care reform debate. *Politics* 8(1):5.

Oh, O.; Kwon, K. H.; and Rao, H. R. 2010. An exploration of social media in extreme events: Rumor theory and twitter during the Haiti earthquake 2010. In *Proc. ICIS'10*, number 231.

Owen, A. B., and Eckles, D. 2012. Bootstrapping data arrays of arbitrary order. *Ann. Appl. Stat.* 6(3):895–927.

Prasad, J. 1935. The psychology of rumor: A study relating to the great Indian earthquake of 1934. *British Journal of Psychology* 26(1):1–15.

Qazvinian, V.; Rosengren, E.; Radev, D. R.; and Mei, Q. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proc. EMNLP'11*, 1589–1599.

Starbird, K., and Palen, L. 2012. (How) will the revolution be retweeted?: Information diffusion and the 2011 Egyptian uprising. In *Proc. CSCW'12*, 7–16.

Ugander, J.; Karrer, B.; Backstrom, L.; and Marlow, C. 2011. The anatomy of the Facebook social graph. Technical report. http://arxiv.org/abs/1111.4503.