Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations

Aniko Hannak
Northeastern University
ancsaaa@ccs.neu.edu

Drew Margolin
Cornell University
dm658@cornell.edu

Brian Keegan
Northeastern University
b.keegan@neu.edu

Ingmar Weber
Qatar Computing Research Institute
iweber@qf.org.qa

Abstract

The prevalence of misinformation within social media and online communities can undermine public security and distract attention from important issues. Fact-checking interventions, in which users cite fact-checking websites such as Snopes.com and FactCheck.org, are a strategy users can employ to refute false claims made by their peers. While laboratory research suggests such interventions are not effective in persuading people to abandon false ideas, little work considers how such interventions are actually deployed in real-world conversations. Using approximately 1,600 interventions observed on Twitter between 2012 and 2013, we examine the contexts and consequences of fact-checking interventions. We focus in particular on the social relationship between the individual who issues the fact-check and the individual whose facts are challenged. Our results indicate that though fact-checking interventions are most commonly issued by strangers, they are more likely to draw user attention and responses when they come from friends. Finally, we discuss implications for designing more effective interventions against misinformation.

"Why you all in my ear?
Talkin' a whole bunch of sh*t that I ain't tryin' to hear
Get back mother**cker you don't know me like that"
— Ludacris, *Get Back* (2005)

Introduction

False rumor and other forms of misinformation are important issues of public concern. Many fear that new communication technologies may be accelerating the spread of misinformation, yet others point out that new technologies also offer the potential for a broader dissemination of facts to counter false claims (Garrett 2011). In previous eras, false information shared within casual conversations could only be confronted after an individual undertook to research the claim on their own, perhaps at great cost or delay. Today, individuals can draw on a wealth of well-sourced and widely recognized information sources including government websites, Wikipedia, and other traditional and non-traditional sources to engage in immediate *fact-checking interventions* against false claims issued in digital spaces.

Previous research suggests such interventions are likely to have minimal effects. Studies consistently show that individuals resist updating beliefs in the face of contradictory facts (Lewandowsky et al. 2012). However, the bulk of research has treated fact-checking as a largely asocial activity — something that professional media can do to for audiences — rather than a social phenomenon that occurs in real-world conversations. In particular, research has not considered how fact-checking might operate differently when when the check comes from a friend, a person with whom the individual has an ongoing relationship and debt of mutual accountability, or someone who has a particularly high or low status within their community. A fact-check is a form of criticism and a potential source of embarrassment. Research on social networks and influence backs up Ludacris's argument in the epigraph that individuals' receptiveness to such critiques are contingent upon the qualities of the social tie over which it takes place (Friedkin and Johnsen 1999; Cialdini and Goldstein 2004). Thus, it may not be surprising that corrections in the absence of such relationships only reinforce their beliefs in misinformation and/or dismiss the competing claim as unworthy of consideration.

In this paper we draw on the fine-grained records of discursive behavior available through Twitter to observe real-world fact-checking interventions in action. In particular, we observe cases where individuals deploy fact-checking websites (Snopes.com, PolitiFact.com, and FactCheck.org) in reply to statements made by others. These fact-checking replies, which we call *snopes*, provide a small but well defined sub-population of conversational fact-checks on the internet. By observing the dynamics around these explicit and recognizable interventions, we address three basic research questions. First, what are the typical social structural contexts of fact-checking interventions ("snopes") as they occur in the real-world? Second, what are the consequences of these snopes on subsequent user behavior? Third, what are the macro-level structural features of the networks formed by these differences in fact-checking behavior over individual relationships?

Our results reveal significant differences between the social status, behavioral consequences, and structural positions of users making ("snopers") and receiving ("snopees")

fact-checks. We find evidence that fact-checking behavior over weaker relationships involves significant differences in users' relative status. Consistent with prior findings, most fact-checks go un-acknowledged by the snopee. However, users who are snoped by "friends" sharing a reciprocated following relationship are over three times more likely to respond afterwards than over other relationship types. Finally, these fact-checking interventions coalesce into a network characterized by highly polarized partisan sniping with fact-checks by friends occurring almost exclusively on the periphery. Recognizing the social contexts and motivations for communicating misinformation can better inform the design of more effective fact-checking strategies.

Background and Theory

Misinformation can have dangerous consequences and yet false rumors are surprisingly stubborn in the face of challenge (Lewandowsky et al. 2012; Nyhan and Reifler 2010). Disturbingly, the presentation of "correct" information often has small effects on individual attitudes and beliefs, limited to particular contexts or specialized circumstances (Garrett 2011; Ecker, Lewandowsky, and Tang 2010). Prior work has repeatedly highlighted the social nature of rumor and misinformation, yet misinformation "correction" studies have, on the other hand, focused almost exclusively on asocial interventions in which individuals receive "corrections" from generic others or abstract institutions.

Sharing misinformation is a social activity, where-in the truth value of claims is often less important than the social and group related functions the claim helps perform (Allport and Postman 1947). Specious assertions can help maintain a group's identity and coherence as symbolic displays of a shared worldview (Foster 2004). Assertions can also help to communicate shared norms and values by implying the kinds of ideas that members of the group ought to hold (Baumeister, Zhang, and Vohs 2004). False information can have practical importance for a group, where speculative ideas can help individuals to collectively make sense of novel or highly uncertain situations (DiFonzo and Bordia 2007; Lewandowsky et al. 2012).

Despite the importance of social motivations in sharing misinformation, research on the effects of the impact of factual corrections has largely ignored this social component. The most commonly tested form of correction involves presenting an individual with facts from an asocial source, such as a generic authority or a large institution such as a news organization. The basic consensus from such research is that misinformation is "sticky." In particular, individuals are unlikely to shift attitudes or beliefs based on misinformation when this information is corrected, particularly if the original, false idea is consistent with a belief or worldview they already hold (Ecker, Lewandowsky, and Tang 2010; Nyhan and Reifler 2010). Even retractions, in which the facts are presented by the same source that provided the original misinformation, do not eliminate belief in the original (false) claim (Lewandowsky et al. 2012).

These findings suggest it is difficult to dislodge false information, and that providing accurate facts will not contribute much to such efforts. Given the highly social nature of misinformation communication, however, it is not surprising that these interventions have not been effective. We seek to investigate a different kind of correction which we term a fact-checking intervention or *snope*. We define a snope as an attempt by an individual to get another individual to pay attention to the facts on a given topic or within a conversation. Such snopes can include overt corrections of misinformation as well as broader attempts to inject factual information into online discussions.

We argue that "social snopes" should be more effective than "asocial snopes" for several reasons. Social networks are important carriers of influence regarding beliefs and attitudes (Friedkin and Johnsen 1999). Mechanisms such as the desire to reduce mental conflict and trust in the judgments of friends will shift individuals attitudes to be more similar to that of their friends, family and colleagues (Cialdini and Goldstein 2004). Snopes can be strong signals of such beliefs and attitudes, particularly on topics which friends may not frequently discuss (Garrett 2011; Garrett, Nisbet, and Lynch 2013). Friends are also likely to share common elements of worldviews which in turn makes snopes from these actors more believable and self-consistent (Weick 1995). Whereas snopes from strangers or generic authorities can be threatening, snopes from friends are more likely to provide the necessary context to integrate the facts within the shared worldview, making it palatable and worthy of consideration (Garrett, Nisbet, and Lynch 2013; Kahan 2010).

Social network structures can also carry normative restrictions on behavior. An individual's behavioral decisions are influenced not only by their beliefs but by their assessment of who and how they will be judged (Tetlock 2002). Individuals that share friends are structurally embedded within groups that can create pressures to conform (Granovetter 1985; Uzzi 1997). A snope from a friend embedded in the same group can signal that a general norm of factual reliability is in effect, or that speculations, whether position-consistent or not, will not be tolerated (Cialdini and Goldstein 2004).

Finally, status and leadership may play important roles in how individuals react to fact-checking interventions. Research on corrections suggests that individuals are more likely to update beliefs when they are corrected by leaders of their group (Lewandowsky et al. 2012). More broadly, popularity and social status within a social network may indicate the relative risks of scrutiny and reputation that individuals face. Leaders may feel the need to maintain credibility. Alternatively, low status members may feel the need to avoid mistakes and having status and reputation fall further.

Social network information has been used successfully in the automatic detection of misinformation on Twitter (Qazvinian et al. 2011). Consistent with our argument, credible tweets tend to come from individuals with a greater number of friends and followers (Castillo, Mendoza, and Poblete 2011), and once detected as a distributor of misinformation, social pressure can lead to Twitter accounts to be shut down (Ratkiewicz et al. 2011). At the same time, evidence suggests users are right to be wary of others' judgments of their credibility in the absence of a shared social

relationship. Users do a poor job of judging information accuracy based on content features alone, relying instead on user-level information such as user name and other contextual information (Morris et al. 2012; Liao et al. 2012). Thus, there is reason for users to expect that asocial snopes will themselves be misguided.

We thus suggest that understanding the social-structural contexts of snopes is critical to assessing and understanding their effectiveness. This motivates three research questions:

RQ 1 *Who snopes whom?* How does the popularity and status of snopers and snopees vary across relationship types? Are there differences in these users' accounts' ages or the latency of fact-checks?

RQ 2 *Do snopes matter?* What are the consequences of different kinds of social snoping? What social-structural features of snopes provoke the attention of snopees?

RQ 3 *Where do snopes happen?* How do snoping relationships vary across different communities? What does this structure reveal about the motivations and consequences of snoping as it occurs in real-world discourse?

Data and Methods

We expect that "snoping" may take many forms in online discourse. In particular, there are likely many cases where the parties involved are aware that one is snoping the other but this meaning is opaque to outsiders. Unfortunately, addressing our research questions requires cases where it is clear to outside observers that snoping is taking place. To obtain such cases we rely on conversational tweets containing links to fact-checking websites. Specifically, using a corpus of tweets collected between January 2012 and August 2013 from the Twitter "gardenhose" public stream, we identified a subset of 3,969 tweets meeting the following criteria.

Conversational The tweets had to be a reply from a "snoper" to a previous tweet made by a "snopee".

Link to a fact-check The tweets had to include a URL that resolved to one of three prominent fact-checking websites in the U.S.: Snopes.com, PolitiFact.com, or FactCheck.org.

We note that this and subsequent filtering steps create a very conservative sample as it excludes many instances of social snoping involving tweets containing no links, links to other websites, or are not recorded as being parts of conversations. This process generated an edgelist of tweets containing a fact-checking URL from snopers replying to or directed at the parent tweet (snopee) of the thread. For the analysis of changes in the snopee's behavior after the fact-checking intervention we collect the history for each of them using the public Twitter API. Concretely, for each snoper or snopee, we get (up to) 5,000 of their followers, (up to) 5,000 of their followees and (up to) their most recent 3,200 tweets. The follower/followee graph was extracted as of January 2014 and historic follower/followee information is not available through the API. Given that only the most recent 3,200 tweets for a given user could be obtained, we had to make sure we have enough history before the fact-checking

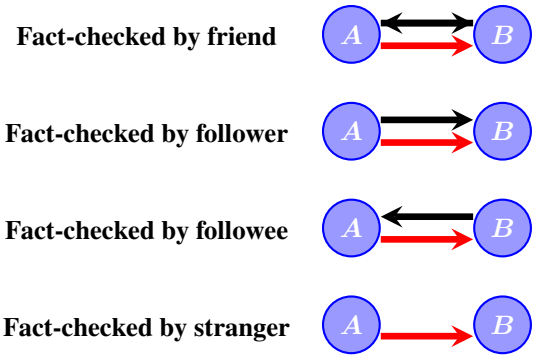

Figure 1: Four dyadic social contexts of fact-checking interventions. For all events where A fact-checks B (in red), B and A can follow each other (friend), A can follow B (follower), B can follow A (followee), or A and B do not follow each other.

intervention. Thus we picked out the users for whom we had historical data starting at least one day before the event. This left us with with 1,614 unique snoping events of which we selected 1,369 snopees who were snoped only once and 1,299 unique snopers who performed these snopes.

For each of the unique snopers and snopees identified above, we extracted the set of users they followed to create a directed following network. We then subsetted this following network to only include the users involved in fact-check interventions as well as the follower edges among these users.

We overlaid edge attributes on top of this subset network based on four distinct dyadic following contexts highlighted in Figure 1. Bob (B) was **snoped by a friend**[1] if Alice (A) fact-checked him and they mutually follow each other. The meaning of "friend" in this context is not meant to convey an offline relationship but simply the fact that they share some degree of mutual accountability by virtue of their reciprocated following relationship. Bob was **snoped by a follower** if Alice, who follows him, also fact-checks him. Unlike with friends, in this case Bob is much less likely to be aware of or concerned with who Alice is. Bob was **snoped by a followee** if he follows Alice and received a fact-check from her. Finally Bob was **snoped by a stranger** if Alice, who does not have a following relationship of any kind, fact-checks him. Intuitively the third scenario (fact-checking a followee) is very unlikely to happen and indeed we have very few data points in this category. We will use these four types of snoping relationships in subsequent analyses.

Direct observation of snoping tweets indicated that while many were used to challenge or correct a statement made by another user, some were intended in other ways, such as to answer a question regarding facts or to make a joke. To determine whether a snoping tweet was challenging the factual basis of the tweet it snoped we used coders from CrowdFlower.com. CrowdFlower is a crowdsourcing platform where paying clients can submit micro tasks, also re-

[1]We deviate from official terminology where a "friend" in Twitter's terminology is a "followee" in our terminology.

ferred to as Human Intelligence Tasks (HIT), and paid contributors/coders can volunteer to work on these tasks. Each task is typically a small set of multiple choice questions and the coders get paid a few dollar cents for completing it.

Coders were instructed to code as a "challenge" any snope in which the snoping tweet attempted to show that facts asserted in the snoped tweet were wrong or that the view expressed by the snopee was based on incorrect facts. For example, a user tweeted:

> ann romney is killing my world series buzz @oshanada http://t.co/LTjL1EKk

This link led to a website that suggested Ann Romney had questioned whether women deserve equal pay. This tweet was then "snoped" by another user who wrote:

> @snopee @otheruser Sorry folks, it's a hoax. The answer is here http://t.co/OmNS7Z19 Yes, we need equality, but facts need to be checked :)

This link led to a refutation of this story on Snopes.com. This pair of tweets was coded as a "challenge."

In contrast, not all links to the websites we identified were necessarily fact-checking interventions. For example, one user wrote:

> @snoper What is the fact checking website you showed us in class again?? It's so hard to keep up with whats the truth or not!

What we would have classified as a "snoper" replied with a link to base PolitiFact.com domain although the intent was clearly not to fact-check this "snopee". While challenges were relatively easy to identify, no clear set of alternative categories was readily apparent.

Relationships between snopers and snopees

Our first research question asks, "who snopes whom?" We begin by identifying the social relationships that typically exist between snoper and snopee. Using the complete list of snopes (rather than the unique snopee subset), we observe 1,290 total snopers, 1,310 snopees, and 36 users who both snoped and were snoped. For each of the four fact-checking relationship types identified in Figure 1, 773 snopes were made by strangers, 442 snopes by friends, 267 snopes by followers, and 51 snopes by followees. Because of their scarcity, we ignore snopes by followees in subsequent analyses. This first descriptive finding indicates that snopes between individuals who are likely to share some sense of mutual respect and accountability are in the minority (roughly 30% = friends), but are not unheard of. Stranger-stranger snopes are the most common form, however.

Structural position

We examined the relationship between snopers and snopes by considering the size and scope of their ego-networks. First, we analyzed the structural position of these users by identifying differences in absolute popularity. Second, we examined the average connectivity of their neighbors to differentiate "celebrities" with many poorly-connected followers from "elites" with fewer but better-connected followers.

In Figure 2, the number of followers for snopers (red) and snopees (green) are plotted across three types of fact-checking interventions. Fact-checks over friend relationships are marked by users' similarity in structural position. Within friend-to-friend snopes, both the snoper and the snopee have similarly-sized audiences. These audiences are also comparatively small. By contrast, fact-checks over follower and stranger relationships are marked by significant (one-way ANOVA, $F = 13.9, p < 0.01$) disparities in structural positions. Snopees in these asocial conditions where they are unlikely to "know their snoper" tend to have tens of thousands of followers on average while snopers are less popular with only hundreds of followers.

We next extend this analysis to the average connectivity of snopers' and snopees' followers as a measure of the user's status as a "celebrity" or an "elite." In Figure 3, the average number of followers' followers are plotted across the three types of fact-checking interventions. As before, fact-checks that occur over friend relationships pair users with similarly-sized audiences. By contrast, fact-checking interventions over following and stranger relationships involve significant asymmetries in users' audience characteristics. In both of these non-friendship cases, snopers' followers have significantly larger followings (one-way ANOVA, $F = 110.6, p < 0.01$) on average than the followers of snopees. In other words, snopers are comparatively "elite": they are followed by users who have more followers themselves than the users who follow snopees.

Finally, we examine the extent to which snopers' and snopees' following networks exhibit clustering. High levels of clustering indicate highly embedded networks in which users' followers also follow each other. In Figure 4, the clustering for snopers and snopees are plotted across the three types of fact-checking interventions. Again, fact-checks by friends and fact-checks by strangers both involve users with levels of clustering. In contrast, fact-checks by followers show a significant difference (one-way ANOVA, $F = 30.8, p < 0.01$): snopers have much more densely-connected local following networks than snopees.

Together these findings present a potential puzzle because snopers tend to have smaller direct audiences but these direct audiences are more tightly connected and have larger indirect audiences than snopees' indirect audiences. This suggests a specific social ecology in which activists embedded within local communities of well-connected actors potentially feel emboldened to call out celebrities.

Temporal relationships

We examine the temporal features surrounding fact-checking interventions to better understand the dynamics of attention to misinformation. Because users who joined Twitter earlier may be motivated to protect the platform from misinformation while users who joined later may lack the skills or shared norms against misinformation, we examine whether there are systematic differences in the ages of snopers' and snopees' accounts. However, we found no significant differences across relationship types between snopers' and snopees' ages.

Because information on Twitter is primarily shared via

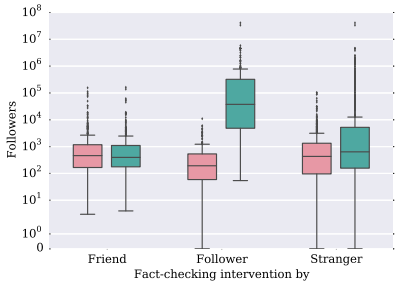

Figure 2: Number of followers.

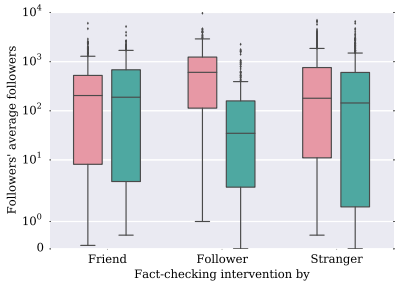

Figure 3: Average followers' followers.

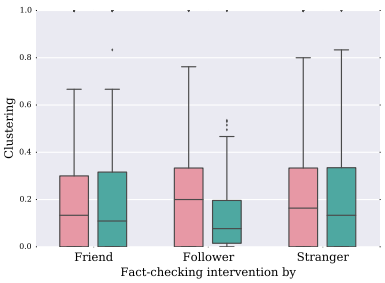

Figure 4: Clustering.

	Mentioned after			
	Yes		No	
Mentioned before	74	(58%)	54	(42%)
Not mentioned before	77	(7%)	1114	(94%)

Table 1: Cross-tabulation of a snopee mentioning the snoper 10 minutes before and after the identified snope.

following relationships, tweets from followees should be seen more quickly and, if misinformed, fact-checked more rapidly. However, there are no significant differences in the temporal lag between snopees' and snopers' messages across relationship types: the bulk of fact-checking interventions occur within an hour of the snopee's statement.

Behavioral consequences of fact-checking

Our second research question asks: How do individuals respond to being snoped? In this section we explore the responses that fact-checking interventions provoke from the snopees to whom they are directed. That is, the "snope" is considered as an experimental treatment provided by the snoper to the snopee. We then examine different qualities of these treatments, specifically, from whom they were issued and whether they "challenged" the respondent and affected their response. Based on the results reported above about the typical time lag between snoping tweets and the tweets to which they reply, we focus our report on snopees' responses over short time windows — 5 to 10 minutes. We have run analysis for up to 24-hour windows following the "snope" and found no substantial changes. All effects in longer time windows are either significant (but weaker) or no longer significant but in the same direction, indicating an expected decay in the effect of the treatment (snope) over time.

Do Snopes Get Recognized?

Our first analysis in this section considers whether snopes get the attention of the individuals to whom they are addressed. Individuals' Twitter streams, particularly amongst highly followed individuals, can be filled with a large number of replies. While the actual cognitive attention extended by snopees to individual tweets in their feed cannot be observed, we can observe how often a snoper is recognized by a snopee in subsequent tweets by considering whether and when the snopee mentions the snoper.

	Mentioned before		Mentioned after	
	Yes	No	Yes	No
Follower	4 (66.7%)	2 (33.3%)	6 (2.9%)	199 (97.1%)
Friend	33 (56.9%)	25 (43.1%)	45 (12.0%)	329 (88.0%)
Stranger	37 (57.8%)	27 (42.2%)	26 (4.2%)	586 (95.8%)

Table 2: Cross-tabulation of previous and subsequent mention of snopers by snopees across relationships types over 10-minute window.

Using this information, we divide our snoper-snopee conversations into two types. The first type are snopes "within ongoing conversation," where the snopee had mentioned the snoper in the minutes prior to the snoping event. The second type are "out-of-the-blue" snopes, where the snopee has not recently mentioned the snoper. As shown in Table 1, there is a substantial difference. The vast majority of snopes occur "out-of-the-blue." Moreover, the distinction between "ongoing" and "out-of-the-blue" snopes is important for whether the snope is subsequently recognized. If a snopee has already mentioned a snoper 10 minutes or less prior to getting snoped, they are likely to mention them after the snope. When the snoper has not been mentioned recently, however, the probability of being mentioned after a snope is significantly smaller (6.5%) ($\chi^2 = 300.60, df = 1, p < .001$). These patterns are robust to at least 24 hrs of conversation, where 2/3 of users mentioned within 1 day prior to the snope are mentioned within a day after the snope, but only 1 in 9 users not mentioned prior are mentioned subsequently.

We next consider how recognition relates to the dyadic relationship between the snoper and the snopee. Table 2 shows the difference between the likelihood of a response to friends as opposed to other relationship types when under the "out-of-the-blue" versus "ongoing conversation" snope conditions. For "ongoing conversations," there is no significant difference. No matter whom the snope comes from, snopees are roughly equally likely to respond. For "out-of-the-blue" snopes, however, there is a substantial difference.

	Unchallenged		Challenged	
Follower	6	(26.1%)	17	(73.9%)
Friend	38	(34.2%)	73	65.8%
Stranger	19	(15.0%)	108	(85.0%)

Table 3: Cross-tabulation of challenged and unchallenged snopes by relationship type.

	Mentioned after			
	Yes		No	
Follower	4	(29%)	10	(71%)
Friend	26	(45%)	32	(55%)
Stranger	20	(27%)	53	(73%)

Table 4: Cross-tabulation of challenged snopes by relationship type for 10-minute window and out-of-the-blue conversations.

The table shows that friends are about 3 times more likely to respond to "out-of-the-blue" snopes from friends than they are from strangers ($\chi^2 = 21.05, df = 1, p < .001$). This finding provides evidence that the ongoing relationship and mutual accountability that are part of a reciprocated following relationship on social media has impacts on how individuals react to snopes. That is, while users tend to ignore "out-of-the-blue" snopes in general, they are significantly more likely to response to "out-of-the-blue" snopes from friends — individuals whom they follow and who follow them — than from non-friend relationships.

One possible explanation for this phenomenon is that the friends are using fact-checking websites differently when replying to other friends. That is, friends may provide references to fact-checking websites as support for their friends point of view, but primarily to challenge or criticize strangers. To examine this possibility, we hand-coded snopee and snoper tweets for all cases where the snopee recognized the snoper within 24 hours after the snoping event. Four coders examined each pair of tweets, and the 261 pairs where at least 3 of the 4 coders agreed were retained for further analysis. Snoper–snopee tweet pairs were coded as either "factual challenges" (76%) or "other" (24%).

Table 3 shows a clear distinction. For friends and (non-mutual) followees, snopes are substantially less likely to be challenges. While the vast majority of snopes from strangers (85%) are challenges, only 66% of those from friends are challenges ($\chi^2 = 11.39, df = 1, p < .001$). This raises the question as to whether friends are recognizing one another's (out-of-the-blue) snopes more frequently because they are less challenging. We thus examine the rate of recognition for challenging snopes.

Table 4 shows the extent to which an individual recognizes challenges when they come from friends versus strangers within a 10 minute time window. Results show that the observed increase in responsiveness to friends' snopes is not explained by these snopes being less challenging, as the bias toward responding to friends over strangers remains even when only challenging snopes are considered ($\chi^2 = 4.31, df = 1, p < .05$).

A related concern is that individuals may simply attend more to friends' messages in general, thus leading to a greater rate of reply for any friend message. If this were true, it would not deny the importance of social relationships in fact-checking, but would suggest that the relevant mechanism is general attentiveness, not responses to the critical content of snopes. We thus consider non-challenging snopes as a base case as in these cases individuals received a reply with a snoping url but were not directly challenged. Though the samples are small ($n = 15$ for non-challenges by strangers; $n = 20$ for non-challenges by friends), we find that individuals reply to both challenges and non-challenges from strangers at virtually the same rate (27%). However, friends reply to 45% of challenges from friends, compared with only 25% of non-challenges from friends ($\chi^2 = 2.44, df = 1, p = .12$). This evidence suggests that the significant differences in rates of reply observed in our larger sample cannot be attributed to friends paying greater attention to replies from friends in general. In fact, though the sample is too small to be conclusive, there is evidence that friends are more likely to respond to challenges from friends than to general replies from friends.

In summary, even though strangers are more likely to snope with a challenge, they are significantly more likely to be ignored. These results provide substantial evidence of the critical role that established social relationships play in the dynamics of misinformation. Though recognition does not necessarily imply persuasion or acceptance, it suggests that the first step toward such persuasion, attention, is already skewed toward friends.

Table 5 shows examples of the tweet streams in which snopers challenged snopees and received a subsequent reply. Though we do not formally model it in this analysis, this discourse is representative of a general pattern in which stranger-stranger snopes contain more hostile or abrupt tones. In each of the first three cases, the snopee rejects the assertion of the snoper, that is, the snope appears to have no effect. Yet the friend-friend snope contains the acknowledgment "I took a look at the article" and then provides evidence to refute the original snope (a link to a website that suggests Snopes itself is biased). By contrast, the replies to strangers simply assert counter-claims without any real justification. The fourth example, a friend–friend snope, shows the idyllic deliberative form which scholars and commentators tend to praise in which the snopee revises his beliefs and publicly retracts his misinformed statement.

Do Snopes Change Discourse?

The preceding section explored whether there were differences in the extent to which snopees acknowledged snopes, and challenges in particular, within different relationships. In this section we examine how snopes change the subsequent discourse of the snopees. Our analyses again are divided into the categories used above: the source of the snope, whether it is recognized, and whether it is a challenge. We then use these as independent predictors of the rate at which the snopee sent tweets after being snoped.

Tweet rate is an important variable to consider because of the possibility that snopes have a "chilling effect" on tweet-

Order	User	Tweet
		Ongoing conversation, Snoped by stranger, Political
Original	Snopee	@Snoper what facts are those? You've been wrong ever since you opened your dumb mouth.
Snope	Snoper	@Snopee Red State Socialism' graphic says GOP-leaning states get lion's share of federal dollars http://t.co/dMGeGZlzNZ TOLD YOU DICKFACE!
Reply	Snopee	@Snoper they block bad bills. Like Obama care, cash for clunkers, and stimulus this is just more bureaucratic BS.
		Ongoing conversation, Snoped by friend, Political
Original	Snopee	@Snoper people have been talking about that for years before OBAMA came into office.
Snope	Snoper	@Snopee but its actually been passed now and Obama signed it http://t.co/Hkp2TztS March 23, 2013 Americans will be REQUIRED to have it!
Reply	Snopee	@Snoper I took a look at the article and "http://t.co/aeLzcNen Rumor has it" is not a credible source. http://t.co/JcyofiZO
		Out-of-the-blue snope, Snoped by stranger, Political
Original	Snopee	Does @SenTedCruz know that #FastandFurious began under George W. Bush?
Snope	Snoper	@snopee @SenTedCruz Check facts B4 running mouth.http://t.co/t7jcBppbLF
Reply	Snopee	@User2 @SenTedCruz Same exact gun-walking program, with a different name. Didn't hear right-wing outrage then.
		Out-of-the-blue snope, Snoped by friend, Non-political
Original	Snopee	Here's some advice Bill Gates recently dished out at a high school speech about 11 things they did not really learn http://t.co/kPgnpdXn
Snope	Snoper	@snopee no, no it isn't... http://t.co/x8m8TxBE
Reply	Snopee	@Snoper; @self no, no it isn't. http://t.co/d6Oochta well spotted Steve, ok it isn't. Some advice from Charles J. Sykes..

Table 5: Four example Twitter conversations coded as challenges containing the original tweet from the snopee, the snoping tweet from the snoper, and the reply tweet from the snopee.

ing, that is, when individuals are challenged or rebuked, they may shy away from further tweeting (at least temporarily). Measuring changes in tweet rate appropriately is difficult, however. By definition, snopees have been active users prior to the snoping event, as it is through the production of discourse that they become candidates for receiving replies. Thus, we know that all tweeters sent at least 1 tweet before being snoped. In a sense, a metronome that sent 1 tweet every hour would appear to reduce its tweet rate following any snopes that responded to it within a short time window.

A conservative way to control for this effect is to consider whether any reduction in tweet rate is greater than 1 tweet per minute of the observed time window. We first test whether tweet rates drop across all cases. We find that rates are significantly lower for time windows of 5 minutes, 10 minutes, and 60 minutes. However, the extent to which they are reduced is not greater than that which may be due to the presence of the snoped tweet. For example, the 95 percent confidence interval in the difference between tweet rates for the 60 minutes prior to the snope versus the 60 minutes after the snope is between .001 and .014 ($t = 2.30, p < .05$). But 1 tweet in 60 minutes is worth .017, so this difference cannot rule out the influence of this "guaranteed" effect.

We next consider whether "challenging" snopes have a significant impact that is distinct from other kinds of snopes. Using a linear model predicting tweet rate from a dummy code for whether the snope is a challenge, we find no significant relationship for any time window (5, 10, 60 or 120 minutes), and the regression explains less than 2% of the variance in tweet rate in each case. We then test for whether the friendship relationship makes a difference. Using a linear model, we estimate the relationship between whether a tweet is a challenge and whether it comes from a friend and find no significant results or improvement in explained variance. This may be due to the fact that, as shown in the analysis of replies, individuals are somewhat motivated to reply to being snoped with a challenge, thus offsetting the possible chilling effect. A more sophisticated model that conditions on an expected number of potentially defensive replies before estimating change in tweet rate may be required but is beyond the scope of the current analysis.

Community structure

Our third research question asks: How do snoping relationships vary across different communities? The preceding analyses were focused on the relationships and interactions between the snoper and the snopee. Here we examine the interactions within the population of users in our dataset aggregated into a large-scale, complex network of snopers and snopees interacting via following relationships as well as snoping communications. We extract and visualize the giant connected component of the follower graph in two steps.

First, we visualize the follower graph in isolation in Figure 5 to illustrate the underlaying social "substrate". Ev-

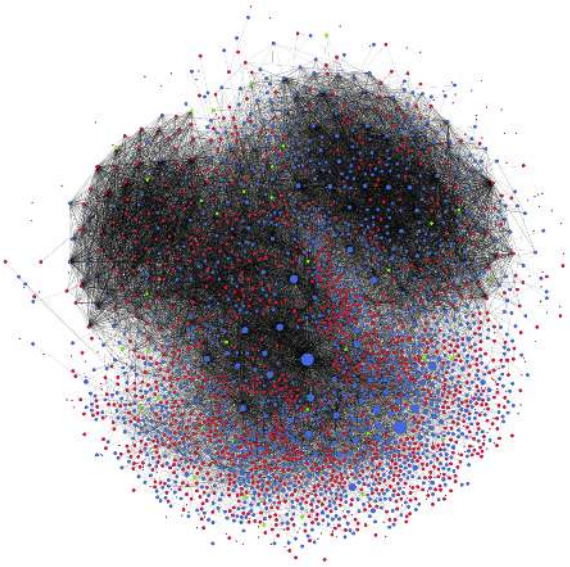

Figure 5: The largest connected component of the follower graph containing 2,636 nodes and 25,618 edges. Snopees are in blue ($n = 1290$), snopers ($n = 1310$) are in red, users who were both snopers and snopees in green ($n = 36$).

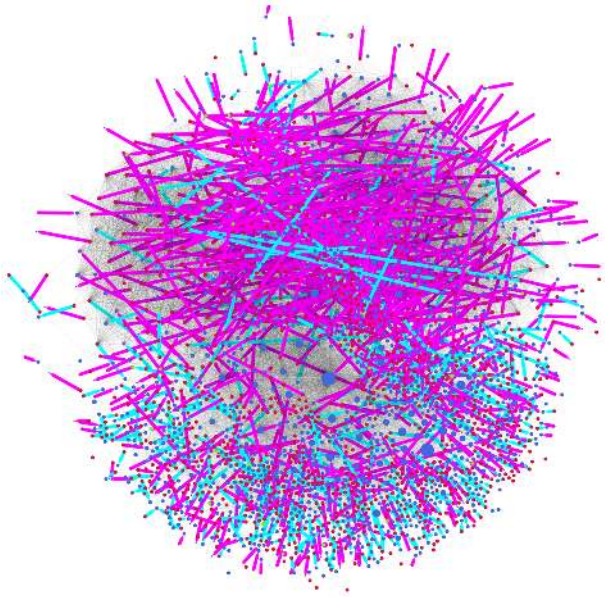

Figure 6: The same graph overlaid with snoping relationships. Pink edges were "snoped by strangers" ($n = 773$) and cyan edges were "snoped by friends" ($n = 442$). "Snoped by followers" ($n = 267$) and "snoped by followees" ($n = 51$) edges are not shown for clarity.

ery node in this network was either involved as a snoper or a snopee from the corpus above. The network is laid out using the ForceAtlas2 spring-embedding algorithm in Gephi by following relationships among users. This follower network exhibits a classic polarization (Adamic and Glance 2005; Conover et al. 2011) between two densely-knit communities. Manual inspection suggest they correspond to U.S. political parties, with conservative political officials and activists (top right) such as `@mittromney` and `@karlrove` and liberal officials and activists (top left) such as `@whitehouse` and `@joebiden`.

Second, we superimpose the snoping relationships from strangers (pink) and from friends (cyan) over this network in Figure 6. "Snoped by followers" and "snoped by followees" are omitted from the visualization for the sake of parsimony. The distribution of these different kinds of snoper–snopee relationships reveal that snoping takes place primarily between these clusters where, by definition, there are fewer friendship relationships. The graph is heavily populated by stranger dyads (pink) while snopes via friend relationships (cyan) are very rare.

This disparity suggests a perverse dynamic that sheds new light on commonly voiced concerns regarding partisan division and bickering. Whereas prior research has established that attention to information is disproportionately focused within political groups (Adamic and Glance 2005), this figure suggests that snoping is explicitly directed outwardly as criticism of individuals that snopers normally do not care to hear. Snopers spiritedly "snipe" opponents' messages despite having no interest in being a member for their messages generally, and perhaps unsurprisingly, these stranger-stranger snopes are subsequently ignored.

By contrast, within-community corrections are rare. This macro-social analysis suggests that fact-checking may play a role not generally foreseen in the discussion of democratic deliberative practice. Snoping tweets go to a unique audience — the snopee and all of the followers of the snoper. However, the polarized nature of the following graph as well as the specific affordances by which Twitter shares reply tweets to only those users who follow both the "replier" and "repliee" suggests few users in snopers' audiences see these tweets. Though not definitive, the results here suggest that snoping is used less as an invitation to others to behave more deliberatively and more as a performance or ritual to signal one's loyalty to a community and willingness to make public declarations against opponents.

There is also a more diffuse third community at the bottom of Figure 6 containing celebrities such as `@rihanna` and `@neiltyson`. Fact-checking interventions within this "popular culture" community have substantially more "snoped by friend" relationships (cyan) than "snoped by stranger" relationships. This may simply be an effect of celebrities reciprocating following relationships unselectively, but it also points to the toxic consequences of polarization and selective exposure. In conjunction with the prior findings about structural position, these users are equals to each other but peripheral to celebrities and elites within the population. This portends a more democratic ideal for deliberative speech and may be opportunities for users to revise their beliefs in or support for misinformation.

Discussion

This research proposed to examine the fact-checking as a social activity. Though exploratory in nature, the results support the conclusion that within real-world conversations, the social attributes of a snope are important to understanding its intent and impact.

Through our examination of the social component of fact-checking, our findings suggest that snopes are often performative acts intended more for the snoper's audience. Nonetheless, good-faith communicative acts actually intended for the snopee's consideration can be effective in prompting a response if it occurs over existing friendship relationships.

Clearest among these findings is the importance of pre-existing relationships. The majority of observed snopes, drawn at random from public discourse, come from strangers — individuals with no existing social relationship between them on Twitter. Fewer than 30% of snopes come from friends. Yet "friendly" snopes are much more likely to get an individual's attention. The difference appears to be explained not by greater attentiveness to friends in general, but by a tendency for individuals to respond to friends' challenges. Whether these responses are acknowledgments of error or attempts at self-defense to save face, it is clear that the existing social relationship between the parties encourages the individual to consider and address the challenge. The importance of ongoing social relations is also highlighted by differences in the way individuals react to snopes within ongoing conversations and snopes that occur "out-of-the-blue." Individuals do not appear to avoid snopers when they present their facts within an ongoing discussion. When snopes are lobbed in from outside such discussions, however, they are very likely to be ignored, especially when sent in by strangers.

Ironically, these stranger–stranger "out-of-the-blue" corrections are most akin to those created in many laboratory fact-checking experiments. Our results suggest that it is unsurprising that these laboratory corrections fail to produce results. However, the culprit is not individuals' resistance to facts, but their resistance to being challenged through asocial channels by faceless strangers. This result is promising for the study of strategies to combat misinformation, however, our data also show that, unfortunately, this least influential kind of snope is also the most typical in real world discourse. Challenges from strangers comprise the most common kind of snope, but they have the lowest chance of garnering a reply. Challenges from friends actually appear to increase discussion, suggesting they may make a contribution to cooperative deliberation.

Status and popularity also appear to play an interesting role in real-world snoping. Individuals with dense networks of highly-connected followers tend to snope popular individuals with sparse networks of poorly-connected followers. Once again, the typical form of snope is the least effective. Bivariate correlations suggest that snopes from high status users tend to be ignored more so than snopes from others, though the effect is weaker when considering friendship.

Are individuals intimidated out of responding to high status individuals, or do they suspect their intentions are not to foster deliberation in the first place? The community structure analysis suggests the latter. Elites appear to snope across communities, attacking the popular members of the groups of which they are not a part. This suggests these activities may be performative rather than deliberative, with elites garnering and maintaining status through snopes that are visible not only to the snopee but their entire set of followers.

Limitations and future work

First, this corpus is only a sample of fact-checking interventions and does not include tweets from before 2012, follower graph information at the time of the intervention itself, or fact-checks using other websites and data sources. Future work might rely on tie-strength observed prior to and after the snoping event (Bak, Kim, and Oh 2012). Second, the fact-checking interventions we identified were English language and predominantly involved content related to the context of U.S. politics. The variance in cultural norms about political deliberations and social media practices across linguistic and national contexts prevents us from making general claims about the effectiveness of socially-mediated fact-checking interventions. Third, the interventions were used as natural experimental treatments, but nevertheless lacked the independence and random assignment to conditions to definitively establish the effectiveness of socially-mediated fact-checking interventions in encouraging individuals to rethink their views and statements. These limitations might be addressed in the future by automated or semi-automated classifications of misinformation and random selection from different types of interventions to test their effectiveness. Finally, more advanced qualitative content analysis and natural language processing methods are required to better the conversational contexts, user dispositions, and topical features that shape decisions to both transmit misinformation and attempt corrections through fact-checks.

Conclusions

New communication technologies have increased the capacity to spread both misinformation and challenges to it. We argued that because the communication of misinformation is a fundamentally social process, the presentation and acknowledgment of facts must also be considered in its social context. Using a corpus of 1,614 conversational tweets referencing well-known fact-checking websites, we identified the social contexts and behavioral consequences of fact-checking interventions. Our results suggest these interventions occur over diverse social contexts and these contexts play a crucial factor in explaining changes in subsequent user behavior. The theoretical framework we have outlined and tested likewise have implications for designing strategies to respond to misinformation in both online and offline contexts.

Acknowledgments

We would like to thank Alan Mislove and Kiran Garimella for assistance with data preparation as well as three anonymous reviewers and the members of the Lazer Lab at Northeastern University for their feedback.

References

Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proc. 3rd Int'l Workshop on Link Discovery*, LinkKDD '05, 36–43. New York, NY, USA: ACM.

Allport, G. W., and Postman, L. 1947. *The psychology of rumor.* Russell & Russell.

Bak, J. Y.; Kim, S.; and Oh, A. 2012. Self-disclosure and relationship strength in Twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 6064.

Baumeister, R. F.; Zhang, L.; and Vohs, K. D. 2004. Gossip as cultural learning. *Review of General Psychology* 8(2):111–121.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on Twitter. In *Proceedings of the 20th international conference on World wide web*, 675684.

Cialdini, R. B., and Goldstein, N. J. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.* 55:591–621.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on Twitter. In *ICWSM*.

DiFonzo, N., and Bordia, P. 2007. *Rumor psychology: Social and organizational approaches.* American Psychological Association.

Ecker, U. K. H.; Lewandowsky, S.; and Tang, D. T. W. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition* 38(8):1087–1100.

Foster, E. K. 2004. Research on gossip: Taxonomy, methods, and future directions. *Review of General Psychology* 8(2):78–99.

Friedkin, N. E., and Johnsen, E. C. 1999. Social influence networks and opinion change. *Advances in Group Processes* 16(1):1–29.

Garrett, R. K.; Nisbet, E. C.; and Lynch, E. K. 2013. Undermining the corrective effects of media-based political fact checking? the role of contextual cues and nave theory: Undermining corrective effects. *Journal of Communication* 63(4):617–637.

Garrett, R. K. 2011. Troubling consequences of online political rumoring. *Human Communication Research* 37(2):255–274.

Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. *American journal of sociology* 481–510.

Kahan, D. 2010. Fixing the communications failure. *Nature* 463(7279):296–297.

Lewandowsky, S.; Ecker, U. K. H.; 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.

Liao, Q. V.; Wagner, C.; Pirolli, P.; and Fu, W.-T. 2012. Understanding experts' and novices' expertise judgment of Twitter users. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, 24612464.

Morris, M. R.; Counts, S.; Roseway, A.; Hoff, A.; and Schwarz, J. 2012. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 441450.

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

Qazvinian, V.; Rosengren, E.; Radev, D. R.; and Mei, Q. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 15891599.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gon\ccalves, B.; Patil, S.; Flammini, A.; and Menczer, F. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, 249252.

Tetlock, P. E. 2002. Social functionalist frameworks for judgment and choice: intuitive politicians, theologians, and prosecutors. *Psychological review* 109(3):451.

Uzzi, B. 1997. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative science quarterly* 35–67.

Weick, K. E. 1995. *Sensemaking in organizations*, volume 3. Sage.