Different Faces of False: The Spread and Curtailment of False Information in the *Black Panther* Twitter Discussion

MATTHEW BABCOCK, DAVID M. BESKOW, and KATHLEEN M. CARLEY, Carnegie Mellon University, Pittsburgh, USA

The task of combating false information online appears daunting, in part due to a public focus on how quickly it can spread and the clamor for automated platform-based interventions. While such concerns can be warranted, threat analysis and intervention design both benefit from a fuller understanding of different types of false information and of the community responses to them. Here, we present a study of the most tweeted about movie ever (*Black Panther*) in which the spread of false information of four different types is compared to the ad hoc Twitter community response. We find that (1) false information tweets played a small part in the overall conversation, (2) community-based debunking and shaming responses to false posts about attacks at theaters overwhelmed such posts by orders of magnitude, (3) as another form of community response, one type of false narrative (Satire) was used to attack another (Fake Attacks), and (4) the four types of false-information tweets differed in the use of hashtags and in the role played by originating users and responding users. Overall, this work helps to illustrate the importance of investigating “on-the-ground” community responses to fake news and other types of digital false information and to inform identification and intervention design and implementation.

CCS Concepts: • **General and reference** → **Empirical studies**; • **Information systems** → **Social networking sites**;

Additional Key Words and Phrases: Social media analytics, dynamic network analysis, social networks, fake news, disinformation

ACM Reference format:

Matthew Babcock, David M. Beskow, and Kathleen M. Carley. 2019. Different Faces of False: The Spread and Curtailment of False Information in the *Black Panther* Twitter Discussion. *J. Data and Information Quality* 11, 4, Article 18 (August 2019), 15 pages.

<https://doi.org/10.1145/3339468>

1 INTRODUCTION

Fake news has become the news. Over the past several years especially, there has been a growing interest in and concern about the spread of fake news and other types of false information through online social media on the part of members of the public, academic researchers, and policy makers. Different groups have expressed worry about the difficulty in constraining false information and the possible impact of it on civil society.

This work was supported in part by the Office of Naval Research (ONR) under grants - N000141812108 and N00014-17-1-2605 / 25323818CMU and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. government.

Authors' addresses: M. Babcock, D. M. Beskow, and K. M. Carley, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213, USA; emails: {mbabcock, dbeskow}@andrew.cmu.edu, kathleen.carley@cs.cmu.edu.

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1936-1955/2019/08-ART18 \$15.00

<https://doi.org/10.1145/3339468>

This concern has led to a rising demand for personal and policy solutions that limit the perceived and real threat of false information while at the same time maintaining the benefits of a free and open online discourse [21]. A summary of possible policy interventions in the March 2018 edition of *Science* distinguishes between two types of interventions: empowering individuals and platform-based detection and intervention [14]. Empowering individuals involves fact-checking resources and educational efforts to help contextualize information and evaluate sources. Platform-based solutions are described as more top-down algorithmic approaches to minimize the amount of people that see the false content. Each of the general approaches have certain advantages and disadvantages (e.g., education creates resilience but can be slow; platforms can act quickly but can also engage in excessive censorship). A more complete characterization of the false information on social media is needed to help stakeholders and decision makers better understand the level of expected impact and the appropriateness, effectiveness, and efficiency of various interventions.

Three current impediments to a more complete characterization by researchers are (1) a focus on the verified/rumor and true/false dichotomies with less attention to the different types of false information, (2) a focus on how false stories spread and are supported with less attention to the response to those stories, and (3) an intervention discussion that appears focused on normative top-down or bottom-up approaches, with less attention paid to the organic real-world reaction of online communities when faced with false information. Not all false information looks or acts the same or is responded to similarly. Analysis on the broad differences between how verified and unverified stories spread such as that conducted by Vosoughi et al. [26] are helpful but don't tackle how different types of false stories interact, whether they are responded to, and what the impact of that response is. Neglecting exploration of the on-the-ground action of online communities may result in interventions that are not needed or lack efficacy. Dismissing differences in types of false information may lead to inefficient intervention design that considers satire or parables as problematic as propaganda.

In this current work, we present an in-depth case study of the false stories in the Twitter conversation about the *Black Panther* movie, which compares the tweets, users, and hashtags related to four different types of false narratives. We seek to answer how the different types of false information were spread and responded to by the Twitter community involved in the #*BlackPanther* discussion. The next section of this article summarizes relevant research literature regarding different types of false narratives and responses to them. The third section details the case study context and our data collection and analysis methods. Then, we present our results comparing the different types of narratives to each other and to the community responses. The last section contains the discussion of our findings and our concluding thoughts.

2 RELATED WORK

Twitter has been studied as a medium through which “fake news” and other forms of false information are disseminated [3, 6, 19, 26]. A focus of the current Twitter-based research has been on describing the differences in the characteristics and diffusion behavior of “true” versus “false” information and on designing tools to assist in discriminating between them. Methods range from investigating linguistic cues and network characteristics [7] to the actions of bot or cyborg actors [19]. In their recent large-scale study of Twitter, Vosoughi and colleagues concluded that false stories spread farther and faster than true stories and that they were more novel and more closely associated with fear, disgust, and surprise-based emotional responses [26]. Some of these effects were found to be more pronounced for stories involving political content, and politically controversial content has been found to enhance the diffusion of information on Twitter in general [16].

Even after settling debates as to whether a story is verifiably false (which is a process with its own difficulties), there is the additional difficulty in that false information comes in many

different flavors, each potentially with its own purpose and activity. Looking at this issue from the perspective of journalism, Rubin et al. discuss three types of fake news that they label as: serious fabrications (tabloids and yellow journalism), large-scale hoaxes (deliberate falsification causing harm), and humorous fakes (satire and parody) [17]. Also from a more communications-oriented view, Tandoc and colleagues divide fake news into satire, parody, fabrication, manipulation, advertising, and propaganda based on the levels of veracity and deception involved with each type [22]. In a general sense, distinctions may be made between false information that is intended or acts to mislead and misinform (propaganda, hoaxes, swindles), false information that is intended or acts to inform (satire potentially, parable, storytelling), and false information that tests or signals understanding or common thinking (jokes). Importantly, Rubin and colleagues also mention that additional work is needed in identifying different types and their interactions with one another and that it can be difficult to distinguish between different types of false narratives. For example, some recent work has suggested that intentionally fake news articles are more closely related in complexity and style to satire articles than to “true” news articles [13]. There are also many recent examples of satire being mistaken for “fake news.”

In terms of Twitter-oriented research that targets specific types of false information, much of the focus has been on using natural language processing to improve machine-learning-based algorithms to detect satire [2, 18]. Less is known regarding the variations in the activity and dynamic network characteristics between satire and other types of false information such as propaganda. Work by Volkova et al. [25] does utilize some network properties in addition to linguistic features and to classify between propaganda, satire, hoaxes, clickbait, and verified news posts. They compare retweet patterns and find that clickbait, hoaxes, and propaganda are retweeted at higher volumes over shorter times compared to base rumors.

Just as the nature of different types of false narratives is important to consider, so too is the nature of the response to those narratives. Such community-based interventions involve going beyond empowering individuals to correctly evaluate false information. It additionally involves having those individuals act to call out, mitigate, or otherwise attempt to control the spread of false information, whether on their own or in concert with others. Community-led efforts can help highlight and correct, and thus perhaps control, the spread of false or misleading information [1, 8]. Glenski et al. have conducted research to classify types of reactions to false and true narratives and to determine how users react between false and true information sources [11]. They find differences in the type and speed of responses with deceptive posts engendering more accepting responses than elaboration-type responses. In other recent work [12], they measure differences in reactions to false and true news sources from bots and humans, finding different reaction rates between the two types of actors.

A related strand of research has focused on the role of different types of actors. This includes research on bots in spreading false information and distorting the information environment [3, 9, 27]. However, it goes further to show how bots, particularly social-influence bots, shape both groups and messages to foster the spread of their messages [4]. The prevalence of bots in social media, and particularly Twitter, is well documented, e.g., one study suggests that 48 million Twitter accounts may be bots [24]. These bots may play a critical role in the spread of false information, as they often link users to the blogs and on-line sites with the full fake news story or other type of false narrative. Exactly how strong this role is is unknown; however, it may be substantial, given that some have concluded that two-thirds of all links shared in Twitter are shared by bots [28]. Other specific actors of importance include news agents and verified accounts, as the former can act to either inform and fact-check or pass along false information, and the latter, due to their status, may have an easier time sharing false information.

Identified distinctions that exist between false narratives may be beneficial for the development of future classification tools and for improved decision making regarding whether to and how to respond to such false information. Using a social cyber-security approach [5] that mixes both computer and social science, we add to the previous work by exploring the interactions and responses of different false narratives in the same overall Twitter conversation and by analyzing the modes and diffusion of the responses to those narratives (quotes vs. retweets vs. replies).

3 DATA COLLECTION AND ANALYSIS

On February 16, 2018, Marvel Studios' *Black Panther* opened to critical and financial success. As of November 2018, it was the third-highest grossing film of all time in the United States. The film tells the story of the Marvel comic book superhero who becomes the king and protector of the hidden and technologically advanced fictional African nation of Wakanda. *Black Panther* was the first movie in the highly successful Marvel Cinematic Universe series to have a predominately African and African-American cast and focus, a fact promoted both by Marvel's parent company, Disney (who released the movie during Black History Month in the United States), and on social media networks.

Surrounding this movie was a social media storm. Much of the activity was on Twitter and Instagram, and according to *Forbes*, *Black Panther* left *Star Wars: The Last Jedi* and other recent blockbusters "in the dust" [10]. As of December 5, 2018, *Black Panther* is the most-tweeted-about movie of all time [29]. Some of those tweets were found to be originating and spreading fabricated and misleading information. The high level of Twitter activity surrounding the release of this movie and the presence of multiple types of false information being spread makes it an ideal case study for examining variation across types of false information within the same context.

In this article, we focus on a total of four types of false or misleading narratives that were shared on Twitter regarding the film:

- *Fake Attacks.* Early showings of the film began the evening of February 15, 2018. On the morning of February 16, it was reported by *Buzzfeed* that there had been a series of Twitter posts claiming a user or their friends or family had been physically attacked attempting to see *Black Panther* [20]. *Buzzfeed* also reported that soon after some of these original posts, other Twitter users had posted responses that provided evidence that images used in the original posts had been taken from other news and entertainment media and called out the original posts as fake stories aimed at stoking racial conflict (most fake-attack posts depicted white people being attacked by black moviegoers, some the opposite) and tarnishing the film's reputation.
- *Satire Attacks.* During the afternoon of February 16, 2018, *Vox* reported that in addition to posts debunking and calling out the original Fake Attack tweets directly, some Twitter users were also mocking the Fake Attack tweets by posting their own versions using either more clearly unrelated photos or additionally unbelievable language [15]. Additional news reports mentioned that some of these satire posts were being treated as if they were examples of the original fake posts.
- *Fake Scenes.* Several tweets posted starting on February 15, 2018, claimed that the movie contained certain scenes that were not actually in the movie. These tweets contained photoshopped images depicting the scenes in a theater and the audience with their fists raised. The content of the rumored scenes varied but several of the scenes had racial overtones (e.g., images depicting white people being treated as slaves). To the authors' knowledge, none of the tweets of this type were reported on by news media.
- *Alt-Right.* Prior to the release of the movie, several articles, social media posts, and YouTube videos made claims that the *Black Panther* movie promotes the alt-right philosophies

encapsulated by isolationism, closed borders, and anti-diversity. Throughout February, including before the movie was released, there were a number of tweets that made this claim. While such issues have been discussed as part of the character's story in the comics, the claims that the movie promotes these policies or "is alt-right" are very misleading, since in the film such policies are questioned and repudiated (and the film ends with a direct denial of the isolationist philosophy). To the authors' knowledge, none of the tweets of this type were reported on by news media, though there were reports about the alt-right trying to co-opt the movie that mention the original articles and YouTube videos.

We collected all tweets containing "#BlackPanther" that were posted from February 8 to February 23, 2018, using Twitter's API to get a starting set of tweets about the movie conversation. For the Fake Attack and Satire Attack type narratives, we collected online news articles using Google with the search terms "Black Panther Fake News" to find published examples of tweets that shared those narratives. From a collection of 15 news articles that mentioned the Fake Attack and Satire Attack stories, we collected a preliminary set of reported tweets. From this preliminary set, we identified key words and phrases (e.g., "black youths," "beaten up," "MAGA hats," "cracker") and then collected additional tweets that contained these phrases as well as the phrase "fake."

We reviewed these results and additionally discovered the presence of Fake Scene and Alt-Right posts. We manually categorized posts into the four types of false narratives: All posts that made a direct claim that *Black Panther* was Alt-Right through words or images were categorized as Alt-Right. All posts that made a claim of a scene occurring in the movie that did not occur were labeled as Fake Scene. To distinguish Satire Attack posts from Fake Attack posts, we based our categorization on whether the tweet used hyperbolic or nonsensical phrases (e.g., "attacked me with a nuke"), used language that started the same as the Fake Attack but ended with a nice interaction ("...and then the youths high-fived us"), and/or used a picture that was hyperbolic, nonsensical, or was of a well-known cultural reference (e.g., nuclear explosion, cartoon characters).

"Origin posts" are the starting points of our analysis and are defined as the non-retweet, non-reply, and non-quote tweets that make a specific claim or statement. In total, we identified 71 Fake Attack, 178 Satire Attack, 8 Fake Scene, and 46 Alt-Right origin posts. For the total of 303 false information origin posts, we identified a total of 278 distinct users who posted them. We additionally collected timeline tweets for these 278 users. The combined dataset contains a total of 5,989,967 distinct tweets (including retweets). We created a subset for each narrative type that included all origin tweets, all retweets of any of the origin tweets, and all responses to the origin tweets (quotes, replies, and retweets of quotes and replies). We also created a subset that contained all other tweets collected (the majority of the #BlackPanther conversation).

We used Twitter's API to determine verified status and a known list of news agency handles to determine which accounts were news agents. We used CMU Bot-Hunter, an integrated ML approach [4], to provide a bot-likelihood score to each user. We compared the different false-narrative types by the magnitude and timing of the origin tweets, response tweets, hashtags used, and user types associated with each. We also compared the hashtag and user networks and explored in depth the most shared false stories and responses to them.

4 RESULTS

4.1.1 Origin Post Frequencies over Time. The pattern of when the origin posts for each type of narrative were created differs by type. Figure 1 shows the frequency of the posting of the origin tweets. The Alt-Right origin tweets were posted throughout the month of February 2018, whereas the other types of origin posts start on February 15 of that month. The Fake Scene posts stop on February 19, and the Fake Attack and Satire Attack posts both end on February 24. The Satire

Fig. 1. Frequency of origin posts over time by type of false narrative.

Table 1. Summary of Tweets by Type of False Narrative

	Fake Attack	Satire Attack	Fake Scene	Alt-Right	All False	All Other
# Total Tweets	99,758	52,952	1,898	66	154,674	5,823,001
# Origin Posts	71	178	8	46	303	1,157,349
% Total: Origin Posts	0.1%	0.3%	0.4%	69.7%	0.2%	19.9%
% Total: Retweets of Origin	1.9%	89.7%	88.1%	28.8%	33.0%	59.0%
% Total: Quotes of Origin	6.5%	5.9%	6.6%	0.0%	6.3%	5.8%
% Total: Replies of Origin	2.5%	1.5%	0.9%	1.5%	2.1%	1.1%
% Total: Retweets of Quotes	88.0%	2.1%	3.0%	0.0%	57.5%	11.4%
% Total: Retweets of Replies	1.0%	0.4%	1.0%	0.0%	0.8%	2.9%

Attack posts that appear to have been responses to the Fake Attack posts begin appearing the same day.

4.1.2 Summary of Tweets. The amount of Twitter activity differs substantially between the four types of false narratives and overall is far less than that associated with the rest of the *Black Panther* conversation. Table 1 compares the total tweets involved in the four false-narrative types against each other and the non-false-narrative-focused conversations. The total amount of tweets discussing or sharing any of the four false-narrative types make up only approximately 2.6% of the total conversation that we were able to collect. While the total number of tweets related to the Fake Attack stories is the largest of the four, Table 1 shows that for this type of narrative most of those tweets are retweeting quotes of the origin posts and not the origin posts themselves. We manually identified that over 95% of the quotes of and replies to Fake Attack origin posts were negative towards those origin posts. On the other extreme, for both Satire Attack and Fake Scene-type narratives, most of the tweets were retweets of the origin posts. The results presented in Table 1 also show that the Alt-Right conversation was mostly made up of the origin posts themselves, with very little retweeting, quoting, or replying activity associated with them.

Additionally, almost all the retweets are sharing only a select few of the origin posts or the quotes and replies that respond to them. Figure 2 is a histogram of retweets received by any of the 303 origin posts and shows that most of the origin posts of each narrative type received no retweets at all. Only 11 origin posts in total (5 Fake Attack, 3 Satire Attack, 3 Fake Scene) received

Fig. 2. Histogram of retweets of origin posts by type of false narrative. Note that the x-axis is on a log-scale and there is one very large Satire Attack outlier.

Fig. 3. Histogram of retweets of quotes (left) and replies (right) by type of false narrative.

more than 50 retweets each, including one Satire Attack origin post that received 46,371 retweets (see the far-right data point in Figure 2). For comparison, 5,182 origin posts not related to the false narratives received over 50 retweets, with the most retweeted post gaining 121,697 retweets.

Only 8 origin posts (4 Fake Attack, 2 Satire Attack, 2 Fake Scene) were responded to by more than 50 quotes each and only 5 origin posts (3 Fake Attack, 2 Satire Attack) were replied to more than 50 times. As with the distribution of origin post retweets, most origin posts were not quoted or replied to or were only quoted or replied to once. There were 2 Fake Attack and 1 Satire Attack posts that had over 1,000 quotes each, and 1 Fake Attack story had over 1,000 replies. Only one reply and no quotes were found to be related to the Alt-Right stories.

It will be helpful to think of each quote or reply as a separate post with its own level of support (measured in retweets). Figure 3 summarizes the distributions of retweets of quotes and replies. In total, 27 of the quotes (24 directed toward Fake Attack posts, 3 toward Satire Attack posts) received more than 50 retweets. All these quotes were supportive of their target post if targeting Satire Attack posts and were not-supportive if targeting Fake Attack posts. Only 1 quote of a Fake Attack

Fig. 4. Retweets over time of second-most-retweeted Fake Attack post and top responses. This is the earliest Fake Attack post in our data. The top Debunking Quote received magnitudes more retweets than the original Fake Attack post, peaking soon before it appeared that the account was suspended (around 12 PM on February 16). The Debunking Quote response also began well before the first article mentioning the Fake Attack post was published.

Fig. 5. Retweets over time of top Fake Attack post and responses (including Top Satire Attack Post). Like the previous case, the top Debunking Quote received magnitudes more retweets than the top Fake Attack post did and was being retweeted before the first article about the top Fake Attack post was published. Based on the similarity of images and the timing of the posts, the top Satire Attack post appears to be a direct satire of the top Fake Attack post. This Satire post also received magnitudes more retweets than the original Fake Attack post. There seems to have been a delay of over 1.5 days from when the top Fake Attack story was posted and when there was a noticeable uptick in retweets of it.

post that received more than 10 retweets (out of 108) was supportive of the target Fake Attack origin post. Only 3 replies to any origin post were found to receive more than 50 retweets (all negatively directed toward Fake Attack posts). Of the replies that received more than 10 retweets (21, of which 19 targeted Fake Attack posts), only 1 was found to support its target.

4.1.3 Interactions between False Narratives and Community Responses. The timing and magnitude of the negative responses to the Fake Attack posts appear to have played a role in overwhelming or more directly halting the spread of those posts. The interaction between different types of false narratives also appears to have played a role. Figures 4, 5, and 6 summarize how the three most successful origin posts were spread over time and how they were responded to.

Fig. 6. Retweets over time of top Fake Scene post and responses. Note that both axes are smaller than the previous examples. In this case, the top Fake Scene story was not overwhelmed by negative responses. The other top Fake Scene posts followed similar patterns.

Table 2. Summary of Hashtags by Type of False Narrative

	Fake Attack	Satire Attack	Fake Scene	Alt-Right	All False	All Other
#Hashtags in Origin Tweets	4	2	2	16	19	78,108
#Hashtags Used in Conversation	151	66	13	16	226	88,592
#Hashtags Unique to Type	49	25	1	1	76	88,442
Cosine similarity w/Fake Attack	-	0.13	0.04	0.08	-	-
Cosine similarity w/Satire Attack	-	-	0.03	0.12	-	-
Cosine similarity w/Fake Scene	-	-	-	0.14	-	-

4.1.4 Summary of Hashtags. Table 2 summarizes the number of hashtags and unique hashtags for each narrative type and the cosine similarity between them. While the origin posts for the Alt-Right story contained more hashtags than the origin posts for the other types, no new hashtags were added to the conversation by reactions to the origin posts unlike with the other groups. The hashtag #fakenews was the second-most-used hashtag in the Fake Attack discussion and the third-most-used in the Satire discussion (#blackpanther was the most used in all conversations). This hashtag was the most used across all narrative types in aggregate. #blackpantherisaltright was the second-most-used even though only those in the Alt-Right conversation used it. The third-most-used hashtag in the Fake Attack conversation was #wakandaforever, which was also the second-most-used in the overall *Black Panther* conversation. Most hashtags in the Fake Scene group came from a single advertising tweet unrelated to *Black Panther*.

4.1.5 Summary of Users. The types of narratives differ in the number of users involved in each conversation and by the types of users involved. Table 3 summarizes the number of users by type of false narrative and by type of user. There were very few news agents or verified accounts in any of the false-narrative conversations.

The density histograms of user bot scores are different between type of conversation as shown in Figure 7(a) and for the origin users alone in Figure 7(b). The density profile of the Fake Attack and Satire Attack users more closely matches that of the users in the larger *Black Panther* discussion. The density profiles for the Alt-Right and Fake Scene both show somewhat of a bimodal

Table 3. Summary of Users by Type of False Narrative

	Fake Attack	Satire Attack	Fake Scene	Alt-Right	All False	All Other
# Total Users	91,073	51,456	1,745	50	135,212	1,771,365
# Origin Posters	66	173	7	35	278	406,744
% Total: Origin Posters	0.1%	0.3%	0.4%	70.0%	0.2%	23.0%
% Total: News Agents	0.001%	0.0%	0.0%	0.0%	0.0007%	0.02%
% Total: Verified	0.1%	0.2%	0.06%	0.0%	0.15%	1.4%

Some users are not unique between false narrative types (Thus, the total number of users in the all false aggregate is not the sum of the users in the four other categories).

Fig. 7. Density histogram of user bot scores by type for all users (a) and origin users (b).

Table 4. Summary of User x User Networks by False Type

	Fake Attack	Satire Attack	Fake Scene	Alt-Right	All False	All Other
Network Density	0.00001	0.00002	0.0006	0.007	0.000008	0.000002
Network Clustering Coef.	0.0005	0.0005	0.011	0.000	0.001	0.022
Top Users: Out-Degree	FA1	SA1*	FS1*	AR1	SA1*	O1
	FA2	SA2*	FS2*	AR2	FS1*	O2
	FA3*	SA3	FS3	AR3	FA1	O3
Top Users: In-Degree	FA4	SA1*	FS4	AR4	FA4	MichelleObama
	FA5	SA4	FS1*	AR5*	SA1*	Lupita_Nyongo
	FA6	SA5	FS5	AR6*	FA5	O4

*Indicates that the user tweeted an origin post codes instead of user names provided unless user is a publicly known celebrity or company.

distribution that result in there being a higher percentage of users with higher bot scores. Comparing the origin user distributions, both the Fake Attack and Satire Attack distributions have shifted slightly to higher bot scores.

4.1.6 User x User Networks. Whether the most central actors both reacted to many others and are reacted to by many others differs between the different conversations. As shown in Table 4, some of the top users in the Satire and Fake Scene conversations were central in both in-degree and out-degree, and thus they were more active in both being retweeted/quoted/replied to and

Fig. 8. User x user interaction for origin users. Nodes with rings represent a user responsible for one of the top 11 posts. Larger nodes have higher total degree centrality.

retweeting/quoting/replying to others. Origin posters appear to be taking part in other parts of the conversations they are a part of except in the case of the Alt-Right. The top users in the main *Black Panther* conversation were either celebrities or art and cultural commenters.

4.1.7 Origin User Networks. Many origin users had interacted with origin users of different types of narratives prior to the release of the film. Figure 8 summarizes the interaction network (retweet, reply, or quote activity) between the users who tweeted a false narrative origin post. On February 15, prior to any origin posts except for Alt-Right posts, several of the users who would go on to post Fake Attack, Satire Attack, and Fake Scene posts already had had interactions with each other (including 6 of the 9 top posters). Nine of the 11 top false-post users were connected to other origin posters by February 16, and at that time they form the backbone of the main component. Only 2 of the users responsible for top posts did not interact with other origin posters at all. The Alt-Right origin posters, while active since the beginning of February, did not interact with each other throughout our time period, though 3 eventually interacted with other types of users.

5 DISCUSSION

A main takeaway from our exploration of the false information in the *Black Panther* Twitter discussion is the continued importance of understanding how false information and the response to it plays out in the reality of digital social media. Not making distinctions between different types of false information and focusing on how that false information spreads without considering the response can lead to misperceptions about the overall impact of false narratives and to inefficient, unnecessary, or ineffective interventions. Our study provides a more nuanced example of a situation where there was false information being shared but, first, in total the false information made up a small amount of the overall discussion; second, different types of false information did not act the same way; and third and perhaps most importantly, the response to the most-spread type of false narrative was overwhelmingly negative, and the negative reactions spread further than the original posts in terms of retweets and positive reactions. These reactions that overwhelmed the negative posts themselves did so in a variety of ways with their own advantages and disadvantages.

5.1.1 Successful Community Responses: Debunking and Satirizing. Our results show that most of the individual false narratives present in the *Black Panther* discussion did not spread to any real extent, with only 11 origin posts making it to 50 retweets (3 of which were Satire posts). In terms of the number of original posts and the amount of positive feedback garnered in the form of retweets, several quotes aimed at debunking and several Satire Attack posts aimed at mocking the top Fake Attack stories overwhelmed the stories that they targeted (see Figures 4 and 5). Also, it is important to note that the first debunking activities occurred on Twitter very soon after their targets. In the case of the top two Fake Attack stories, while mass media attention may have played some part in the large response, the Twitter community response was already occurring before articles mentioning their targets were published. In fact, many of the news articles spread the news about the fake news along with the news of the responses, resulting in many non-Twitter users' first exposure to the narratives as already debunked ones.

The spread and apparent curtailment of false information in the *Black Panther* case gives flesh to some of the theoretical models of how rumors are controlled. Tripathy and colleagues [23] conducted modeling that suggested locating "beacon" agents in the network to spread "anti-rumors" and respond to rumors faster than a central authority could assist in slowing the spread of rumors. Twitter users who posted debunking quotes or satire—some of whom existing in the User x User network with posters of Fake Attacks prior to the event—may have acted as these "beacon" agents. We don't have a perfect counterfactual, but the fact that the top Fake Scene post did not have popular debunking quotes or satire aimed at it directly and ended up being retweeted more than responded to suggests that the debunkers and satirists who did act in the Fake Attack discussion may have helped curb the spread of Fake Attack stories. Additional suggestive evidence is the sharp cutoff of the response to the first Fake Attack story (Figure 4) that occurred because the origin account was suspended (included in many of the negative quotes and replies were exhortations for users to report the Fake Attack users to Twitter).

While community response towards the most popular Fake Attack stories appeared to curtail them, the different strategies used had different advantages and disadvantages. The top Satire Attack post, for example, seems to have stayed in the conversation longer than the top Debunking Quote even though the quote received many more retweets. This may be due to the more meme-like nature of the satire response. Also, satirical strategies in general can be problematic if the satire is taken as the real thing, which, based on some of the responses to the satire posts, appears to have occurred. Related to another response choice users have, negative replies to the top Fake Attack stories were shared just as widely, if not more, than the original posts, but none of the replies reached the level of the debunking quotes. Replies and quotes also reached different audiences: an origin poster may more readily see and respond to a reply than to a quote.

It is also important to note that, except for the Alt-Right posts that play off established YouTube videos, the creation of the Fake Attack and Fake Scene narratives appears to have been an ad hoc community process without a larger organization or group behind it. Whether satire or debunking, or the shaming/mockery that can be part of those strategies, will be successful at curtailing different narratives in contexts with more coordinated disinformation campaigns is a question for future work.

5.1.2 Variations in False Narratives. In addition to the differences between reactions to the most popular false narratives, our results show differences among the four narrative types in the role users play, the amount and timing of activity, the use of hashtags, and the possible presence of bot-like accounts. While limited to a specific context, these differences are helpful in pointing out the lack of a one-size-fits-all definition of false information based not on what the goal is, but rather on what their attributes are.

In general, the Alt-Right type of false narrative appears to be the most distinct from the other three. The nature of this type of story is different from the others in that it is based on a claim that the movie represents something, rather than a claim that a certain event (attack or scene) occurred. Tweets claiming that *Black Panther* is Alt-Right or pro-Alt-Right were posted throughout the month of February both before and after the film's release, and the conversation related to such tweets was limited to a much smaller set of users, most of whom were original posters. These are both important aspects to consider when thinking about responses to this kind of false or misleading information. For example, whereas a burst of community response in calling out and mocking may have been successful on limiting the Fake Attack stories, such an acute response may not be appropriate for a slower-burn type of dissemination of misleading information among a smaller group as in the Alt-Right case.

In terms of hashtag use, we find that the Alt-Right origin posters used more hashtags in the origin posts, but no new hashtags were introduced into the conversation; whereas for the other three types, the origin posts only had a few hashtags but the overall conversations added many more. One of the top hashtags most connected to #blackpanther in the main *Black Panther* conversation, #whatblackpanthermeanstome, was used by the Alt-Right origin posters but seemingly to no effect. This hashtag was originally started by the director of online marketing for *BlackGirlNerds* and mostly shared by African-American Twitter community members expressing their appreciation of and hopes for the film. Therefore, it appears that two different communities used the hashtag in potential opposition to each other with the Alt-Right posters unsuccessfully attempting a hashtag hijacking. This suggests that hashtag hijacking, and possibly other types of threadjacking, may not work when the original use of the hashtag is strongly supported by the community that first spread it.

As mentioned in the Related Work section, how bot-like accounts assist in the spread of false information is still under study. Our present results indicate possible differences in the presence of bot-like accounts, depending on the type of false information being promoted. The bot-score distributions for the different conversations (Figure 7) show that the distributions for the users in the Fake Attack and Satire Attack more closely resemble the distribution in the main *Black Panther* conversation, suggesting that for the most popular false narratives, the overall bot-like activity may not be that different from conversations without false narratives. That the Alt-Right and Fake Scene discussions may have had a higher proportion of accounts with higher bot scores does not seem to have increased the presence of those narratives in the wider conversation.

It should also be noted that while different, all of the false-narrative types in this case study are meme-like in nature: within each type (and between Fake Attack and Satire Attack types) the origin posts are very similar with slight variations, are shared by multiple users and in many cases appear to be aware of other versions, meeting many of the criteria outlined by Shifman in *Memes in Digital Culture* [30]. Again, the Alt-Right type is the most different as those users mostly used images that were the same (less "mutations").

6 CONCLUSIONS

If all false narratives are spread the same way at panic-inducing speeds with automated accounts and are not naturally adequately responded to, then the desire for and implementation of a centralized and automated platform-based intervention may be appropriate. But, as our current work provides a good example of, false narratives are not all the same, and communities do react in ways that can be beneficial to curtailing such narratives. Understanding the nuances matters to design and implement interventions efficiently and effectively, or to decide if they are needed at all, especially in complex socio-technological spaces trying to balance safety and free expression. Our

work, while only focused on one specific conversation, hopefully assists in pointing out additional possibilities to be considered when classifying the spread of false information and responses to it.

Additionally, this work suggests that the nature and actions of the community at which the false narratives are aimed is critical in determining the extent to which such narratives are spread, whether they are countered, and the length of time that they persist. We see two main effects, both of which should be examined in more detail. First, we find that resilient communities can respond using debunking and satire narratives that overwhelm and potentially shut down fake-news-type narratives. Being able to rapidly form a networked response is key. Second, we see that the size and possibly the network structure of the relevant communities (both those generating and those targeted by the false information) are important. This research suggests that if the false information originates in an ongoing community, which exists both offline and online, then it may persist for a longer period (e.g., the continual presence of the Alt-Right narrative and the short duration of the Fake Scene stories). It also suggests that if the false information targets a large cohesive pre-existing community, then it may require careful message construction to have an impact (e.g., the failure of the Alt-Right hashtag hijacking and the success of some of the Satire posts).

Additional research is required to determine how widespread and generalizable our findings from this case study are. Moving forward, additional research and analysis of interest include what makes community responses such as the ones we described possible or more effective (where do such responses work and where do they not) and what are the structures and networks that facilitate the identification of false information and the effective countering of it using real-time digital community-based action. Future research also needs to consider how the overall structure of the communities and the network position of the senders and responders impact the spread of fake news and other types of false information.

REFERENCES

- [1] Ahmer Arif, John J. Robinson, Stephanie A. Stanek, Elodie S. Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A closer look at the self-correcting crowd: Examining corrections in online rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 155–168. DOI : <https://doi.org/10.1145/2998181.2998294>
- [2] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. Do we criticise (and laugh) in the same way? Automatic detection of multi-lingual satirical news in Twitter. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*. 7.
- [3] Matthew Benigni, Kenneth Joseph, and Kathleen M. Carley. 2017. Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PloS One* 12, 12 (2017).
- [4] David M. Beskow and Kathleen M. Carley. Bot-hunter: A tiered approach to detecting & characterizing automated activity on Twitter. In *SBP-BRIMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. 8.
- [5] Kathleen M. Carley, Guido Cervone, Nitin Agarwal, and Liu, Huan. 2018. Social cyber-security. In *Proceedings of the International Conference on Social, Cultural, and Behavioral Modeling (SBP-BRIMS'18)*.
- [6] P. R. Chamberlain. 2010. Twitter as a vector for disinformation. *Journal of Information Warfare* 9, 1 (2010), 11–17.
- [7] Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news: Automatic deception detection: Methods for finding fake news. *Proc. Assoc. Inf. Sci. Technol.* 52, 1 (2015), 1–4. DOI : <https://doi.org/10.1002/pra2.2015.145052010082>
- [8] Dharma Dailey and Kate Starbird. 2014. Visible skepticism: Community vetting after Hurricane Irene. In *Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM'14)*.
- [9] John P. Dickerson, Vadim Kagan, and V. S. Subrahmanian. 2014. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 620–627.
- [10] Adrienne Gibbs. 2018. “Black Panther” wins Twitter and Instagram, leaves “Star Wars” in the dust. *Forbes*. Retrieved from: <https://www.forbes.com/sites/adriennegibbs/2018/02/23/black-panther-wins-twitter-instagram-thor-star-wars/>.

- [11] Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018. Identifying and understanding user reactions to deceptive and trusted social news sources. Retrieved from: <http://arxiv.org/abs/1805.12032>.
- [12] Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018. How humans versus bots react to deceptive and trusted news sources: A case study of active users. Retrieved from: <http://arxiv.org/abs/1807.05327>.
- [13] Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Retrieved from: *ArXiv Prepr. ArXiv170309398*.
- [14] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, and David Rothschild. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
- [15] Aja Romano. 2018. Racist trolls are saying Black Panther fans attacked them. They’re lying. *Vox*. Retrieved from: <https://www.vox.com/culture/2018/2/16/17020230/black-panther-movie-theater-attacks-fake-trolls>.
- [16] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the International Conference on World Wide Web*. 695–704.
- [17] Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. 2015. Deception detection for news: Three types of fakes: Deception detection for News: Three types of fakes. *Proc. Assoc. Inf. Sci. Technol.* 52, 1 (2015), 1–4. DOI: <https://doi.org/10.1002/pr2.2015.145052010083>
- [18] María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde, Miguel Ángel Rodríguez-García, Rafael Valencia-García, and Giner Alor-Hernández. 2017. Automatic detection of satire in Twitter: A psycholinguistic-based approach. *Knowl.-based Syst.* 128, (July 2017), 20–33. DOI: <https://doi.org/10.1016/j.knosys.2017.04.009>
- [19] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. 2017. The spread of fake news by social bots. Retrieved from: *ArXiv Prepr. ArXiv170707592*.
- [20] Craig Silverman. 2018. Trolls are posting fake claims of being assaulted at showings of “Black Panther”. *BuzzFeed*. Retrieved from: <https://www.buzzfeed.com/craigsilverman/trolls-are-posting-fake-claims-of-being-assaulted-at>.
- [21] Damian Tambini. 2017. Fake news: public policy responses. *LSE Media Policy Project Series (Media Policy Brief 20)*, Damian Tambini and Emma Goodman (Eds.). The London School of Economics and Political Science, London, UK.
- [22] Edson C. Tanoc Jr, Zheng Wei Lim, and Richard Ling. 2017. Defining “fake news” A typology of scholarly definitions. *Digital Journalism* 6, 2 (2017), 137–153.
- [23] Rudra M. Tripathy, Amitabha Bagchi, and Sameep Mehta. 2010. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM’10)*, 1817. DOI: <https://doi.org/10.1145/1871437.1871737>
- [24] Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: detection, estimation, and characterization. Retrieved from: <http://arxiv.org/abs/1703.03107>.
- [25] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 647–653. DOI: <https://doi.org/10.18653/v1/P17-2102>
- [26] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
- [27] Alex Hai Wang. 2010. Detecting spam bots in online social networking sites: A machine learning approach. In *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, Berlin, Heidelberg, 335–342.
- [28] Stefan Wojcik. 2018. 5 things to know about bots on Twitter. *Pew Research Center*. Retrieved from <http://www.pewresearch.org/fact-tank/2018/04/09/5-things-to-know-about-bots-on-twitter/>.
- [29] “Black Panther.” 2018. “Saturday Night Live” among most tweeted entertainment of 2018. *EW.com*. Retrieved from: <https://ew.com/movies/2018/12/05/black-panther-saturday-night-live-most-tweeted-entertainment-2018/>.

Received May 2018; revised May 2019; accepted June 2019