Pretending Positive, Pushing False: Comparing Captain Marvel Misinformation Campaigns

Matthew Babcock, Ramon Villa-Cox, and Kathleen M. Carley

Abstract It has become apparent that high-profile Twitter conversations are polluted by individuals and organizations spreading misinformation on the social media site. In less directly political or emergency-related conversations there remain questions about how organized misinformation campaigns occur and are responded to. To inform work in this area, we present a case study that compares how different versions of a misinformation campaign originated and spread its message related to the Marvel Studios' movie Captain Marvel. We found that the misinformation campaign that more indirectly promoted its thesis by obscuring it in a positive campaign, #AlitaChallenge, was more widely shared and more widely responded to. Through the use of both Twitter topic group analysis and Twitter-Youtube networks, we also show that all of the campaigns originated in similar communities. This informs future work focused on the cross-platform and cross-network nature of these conversations with an eye toward how that may improve our ability to classify the intent and effect of various campaigns.

Keywords Misinformation · Network analysis · Social media · Topic groups · Cross-platform

1   Introduction

Inaccurate and misleading information on social media can be spread by a variety of actors and for a variety of purposes. In potentially more immediately higher stakes Twitter conversations such as those focused on natural disasters or national elections, it is taken for granted that pressure groups, news agencies, and other larger and well-coordinated organizations or teams are part of the conversation and may spread misinformation. In less directly political or emergency-related conversations,

M. Babcock (✉) · R. Villa-Cox · K. M. Carley
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: rvillaco@andrew.cmu.edu; kathleen.carley@cs.cmu.edu

there remain questions of how such organized misinformation campaigns occur and how they shape discourse. For example, the Marvel Cinematic Universe comic book movies are very popular and the most financially successful film franchise to date. As the Marvel movies have expanded their casts and focused on more diverse storytelling and storytellers, they have become flashpoints in the United States for online expressions of underlying cultural debates (both those debated in good faith and those that are not). This trend appears to be part of a wider one in the comic book communities on social media and within comics' news and politics sites [20].

Social media platforms can be places where honest cultural and political debates occur. However, because of the presence of misinformation and propaganda, many such conversations are derailed, warped, or otherwise "polluted". Conversations that tend to polarize the community often carry misinformation [21], and such polluted content serves the purpose of the influencer or the creator of misinformation. Therefore, further investigation into the use of misinformation on Twitter and the groups that push it in the service of cultural debates may be useful in informing community, individual, and government responses for preserving healthy discussion and debate online. To inform future work in this area, we conducted a case study to explore a set of these issues as they appear in the Captain Marvel movie Twitter discussion.

Captain Marvel, Marvel Studio's first female lead-focused superhero movie, was released on March 8, 2019. Much of the Twitter discussion of the movie was standard comic book movie corporate and fan material. However, in the runup to the release of the movie there was also a significant amount of contentious discussion centered on whether the movie should be supported and/or whether illegitimate efforts were being made to support or attack it. The underlying core of the contentious discussions was related to recurring debates over diversity and inclusion in mass media in general and in comics in particular [20]. There were several different types of false or misleading information that was shared as part of these discussions. One of the most prominent examples was the claim that the lead actor, Brie Larson, had said that she did not want white males to see the movie, which was false. This and similar claims about Larson and Marvel were promoted in multiple ways and were the basis for campaigns calling for a boycott of the movie, either directly (#BoycottCaptainMarvel) or indirectly (by promoting an alternative female-led action movie, Alita: Battle Angel through a hijacked hashtag, #AlitaChallenge).[1]

As these campaigns were found to have somewhat similar community foundations but experienced differing levels of success, they make for a good case

---

[1]#AlitaChallenge first appeared on Twitter to support interest in the Alita movie a year before both the Alita: Battle Angel (Feb 2019) and Captain Marvel (March 2019) movies were released. Up and through the release of Alita the hashtag had minimal use. On March 4, 2019 a politically right-wing celebrity/conspiracy theorist used the hashtag to promote their version of the Alita Challenge (i.e. go spend money on Alita instead of Captain Marvel during the latter's opening weekend) and the hashtag soared in popularity. As Twitter users pointed out, this appears to have been disingenuous of the hijacker as prior tweets demonstrate their lack of interest in Alita as a movie.

study to explore the impact of organization and narrative framing on the spread of misinformation. Moreover, unlike what was observed on previous work on false information in Twitter comic book movie discussions [2], the sources of misinformation in this campaign were more well-known organized groups and/or more established communities providing an additional avenue of exploration. In this work, we compare the two misinformation-fueled boycott campaigns through examination of their origins, the actors involved, and their diffusion over time.

2 Related Work

The study of the spread of misinformation on platforms such as Twitter is a large and growing area of work that can be seen as part of a new emphasis on social cyber security [7, 11]. For example, much of that research has been concerned with who is spreading disinformation, and pointing to the role of state actors [18, 19], trolls [23] and bots [13, 27]. Other research has focused on how the media is manipulated [3, 17]. Still others have focused on impact of spreading disinformation on social processes such as democracy [4], group evolution [3], crowd manipulation [15], and political polarization [26]. There are also efforts aimed at collecting these different types of information maneuvers (e.g. bridging groups, distracting from the main point, etc) that were employed by misinformation campaigns into larger frameworks [6].

Most of the research on information diffusion in social media has focused on retweeting. However, there are many reasons why people retweet [16] and retweeting is only one potential reaction to information found on Twitter. Examining simple retweet totals is static, and doing so does not provide insight into the entire diffusion path [1]. Identifying the full temporal path of information diffusion in social media is complex but necessary for a fuller understanding [22]. Moreover, people often use other mechanisms such as quoting to resend messages [8]. In doing so, they can change the context of the original tweet, as when quotes are used to call out and attack the bad behavior of the original tweeter [2]. Examining replies to tweets and the support of those replies also assists in creating a more complete picture of the discourse path. In contrast to earlier approaches, we consider the full temporal diffusion of misinformation and responses, and the roles of diverse types of senders including celebrities, news agencies and bots.

Many characteristics of message content may impact information diffusion [12]. For example, moral-emotional language leads messages to diffuse more within groups of like-minded individuals [9]. Emotionally charged messages are more likely to be retweeted [24] as are those that use hashtags and urls [25]. The use of urls shows the cross-platform nature of misinformation on social media with many urls pointing to other sites such as YouTube. Research only focusing on one ecosystem may mischaracterize both the who and how of misinformation spread. While our focus is on Twitter activity in the present work, we also use Twitter-Youtube connections to better understand the communities involved.

3   Data Description and Methods

From February 15 to March 15, 2019, we used Twitter's API to collect tweets for our analysis. Our goals and methods for tweet collection were as follows:

1. Compare the origination, spread, and response to the two main campaigns to push the misinformation campaign on interest. To do this we collected all non-reply/non-retweet origin tweets that used #BoycottCaptainMarvel and #AlitaChallenge during our period of interest and collect all quotes, replies and retweets of these origins.
2. Explore the conversation around these two campaigns that did not use the two main hashtags. Compare the hashtag-based and non-hashtag-based conversation. To do this we collected all non-reply/non-retweet origin tweets that used "Alita" along with one of a set of keywords used in the contentious comic-book Twitter discussions (e.g. "SJW", "Feminazi") and collect all quotes, replies and retweets of these origins. We labeled this as the "Charged Alita" conversation.
3. Characterize which communities users who spread or responded to the misinformation were from. To do this we collected the Twitter timelines of the central users and all non-reply/non-retweet origin tweets that provide information about the general Captain Marvel movie conversation (using the keywords #CaptainMarvel, Captain Marvel, Brie Larson), and all quotes, replies and retweets of these origins. To use cross-platform information to understand community structure we collected author, subject, and viewership information for all YouTube video URLs shared through Twitter.

For goals 1 and 3 we are relatively confident that our method allowed us to collect that vast majority if not the entirety of the tweets we aimed for due to our intentions to look at very specific hashtags and to obtain a general sense of where such hashtags were used compared to the most general conversation. For goal 2, while we able to collect a large enough sample of tweets related to the campaign, it is probable that some discussions of the Alita Challenge took place on Twitter using keywords we did not search for, and therefore are not part of our analysis.

We rehydrated any available target of a reply that was not originally captured in the first collection. This allowed us to capture at least the first level interaction within the relevant conversations. In total, we collected approximately 11 million tweets. We used a CASOS developed machine-learning tool [14] to classify the twitter users in our data as celebrities, news agencies, company accounts, and regular users. We used CMU BotHunter [5] to assign bot-scores to each account in our data set. We used the python wordcloud library to generate word clouds for the different groups we found based on YouTube video names.

We used ORA [10], a dynamic network analysis tool, to create and visualize Topic Groups and Louvain clustering that was used to explore the Twitter communities that shared and responded to the misinformation campaigns investigated. Topic Groups are constructed by using Louvain clustering on the intersection network between the Twitter user x Twitter user (all communication) network and the Twitter

user x concept network. The resulting Topic Groups thus provide an estimation of which Twitter users in a data set are communicating with each other about the same issues.

4 Results

4.1 Diffusion on Twitter

We compared the diffusion of the original boycott campaign tweets and responses to them over time. As shown in Fig. 1, direct calls to boycott the Captain Marvel movie started more than a month before its official release on March 8, 2019, without gaining much traction on Twitter except for a day or two before the movie's release (green line). In contrast, during the same period there was an increase of discussion of "Alita" using harsh "culture war" phrases aimed at the *Captain Marvel* movie (Charged Alita). Most striking was the relatively rapid spread of #AlitaChallenge after March 4 (i.e. after the new use of it to attack Captain Marvel promoted by politically right-wing commentators). Overall, most support for these campaigns occurred between March 4 and March 12. Figure 1 also shows that most of the support for responses to the original campaign tweets occurred after the movie was released and based on visual inspection the majority of this support was for responses critical of the various boycott campaigns. The #AlitaChallenge campaign, while being the most directly supported, was also the most widely criticized, followed by responses to the Charged Alita conversations. It is noteworthy that the spike in negative responses coincided with the end of the majority of the support
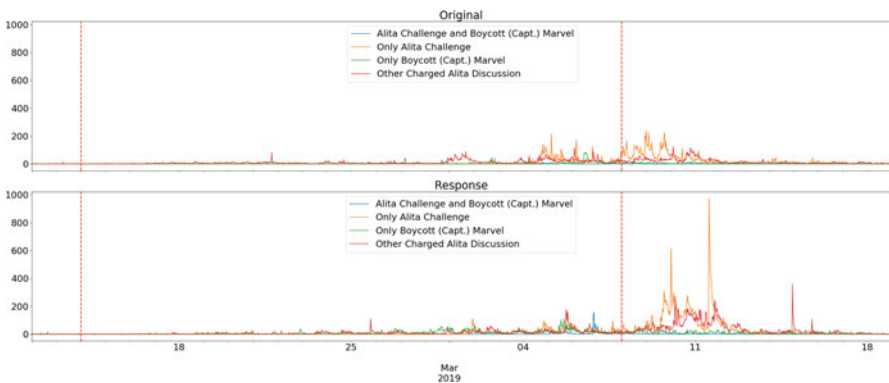
Fig. 1 Diffusion of boycott/anti-Captain Marvel Alita campaigns. Retweets per 30 min window of originals (top) and responses (bottom) for tweets that shared (1) both #AlitaChallenge and #BoycottCaptainMarvel, (2) only #AlitaChallenge, (3) only #BoycottCaptainMarvel, and (4) tweets with "Alita" and harsh words but no #AlitaChallenge. Vertical dashed lines are the opening of Alita: Battle Angel (leftmost) and Captain Marvel (rightmost)

for #AlitaChallenge or Charged Alita tweets. This behavior has been observed on other similar events with contentious messages and debunked rumors [2]. Tweets that directly pushed #BoycottCaptainMarvel or pushed both sets of hashtags were not responded to as great an extent.

## *4.2 Originating and Responding Twitter Communities*

In order to explore which Twitter communities in the overall Captain Marvel conversation the #AlitaChallenge and #BoycottCaptainMarvel campaigns originated and spread to, we calculated the Topic Groups and constructed the Topic Group x Twitter user network. Prior to calculating the Topic Groups, we removed both main hashtags so that the resulting groups would not be based on them as inputs.

Figure 2 shows that a significant majority of origin tweets came from users who were found to be in topic group 8 or 13 and that origin tweets for all four campaign types were found in the same topic groups. Topic Group 8 includes the account that original highjacked #AlitaChallenge as well as accounts and concepts associated with Comicsgate controversies. Topic Group 13 appears to a group more focused on comic book movies in general. Figure 2 also shows that the responses to the different boycott campaigns were more spread out among several more topic groups.

In addition to using twitter actions to communicate (quotes, replies, and retweets), URLs were also shared and helpful in exploring whether the four campaigns were being shared in the same communities. Using the YouTube data mentioned above, we constructed an Agent x Agent network where the nodes are
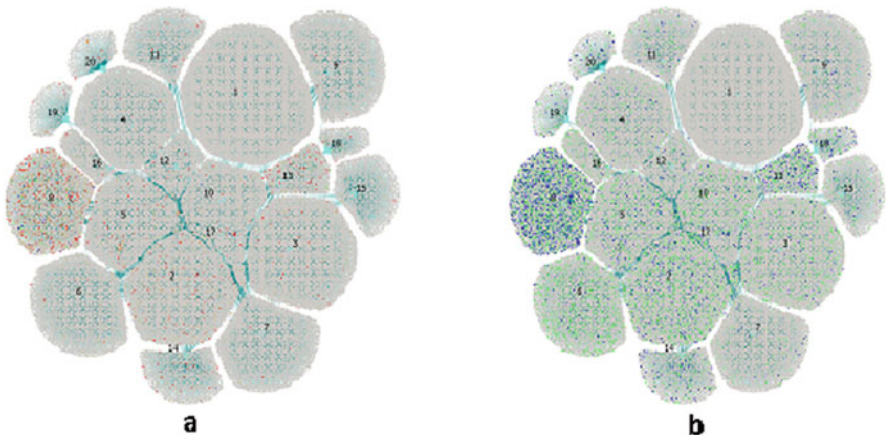
Fig. 2 Twitter Topic Groups Network. Numbers and light blue squares represent Topic Groups. Colored round nodes in Fig. 2a represent origin tweets of the four different types. Colored nodes in Fig. 2b represent the origin tweets and retweets of origin tweets in blue and replies and retweets of replies in green
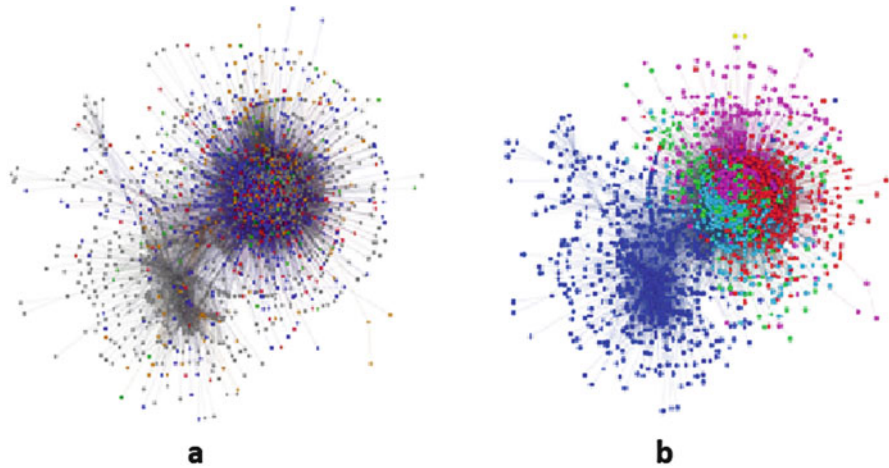
Fig. 3 Twitter user x Twitter user shared Youtube author network. Link weights at least 2. In Fig. 2a, the four colors represent the four types of boycott-related campaigns. Figure 2b is colored by louvain grouping

Twitter user accounts and the links represent the number of shared YouTube authors. For example, if Twitter user A tweets out YouTube videos by author X and Y over the course of our collection period and Twitter user B tweets out videos by X but not Y then the link between A and B is 1 (because they both shared videos by author X). As links with weights of 1 may not be that indicative of broader community (as it could signal only one shared video in addition to one shared author), we examined the subset of this network were the minimum link weight was 2 shared authors as shown in Fig. 2a, b:

Figure 3 shows that there is one main dense cluster of users and one smaller dense cluster. The main cluster was found by inspection to be Twitter accounts mostly involved with anti-Captain Marvel, anti-Marvel, right-wing politics, and "Red Pill" and Comicsgate-like misogyny or anti-diveristy sentiment. The smaller cluster (which is a separate louvain group as shown in Fig. 3b) appears to be Twitter users interested in comic book and other movies without an expressed political/cultural bent. Figure 3a shows that all four boycott related campaigns had origins and support from connected users in the main cluster (note that this cluster appears to have exisited prior to the release of Captain Marvel). It should also be pointed out that the Twitter user who began the hijacking of #AlitaChallenge is not central to the larger cluster though they are connected to the central cluster.

To further help identify the sentiment shared by each sub-cluster, we created word clouds using the titles of the YouTube videos shared by members of those groups as input. The results are shown in Fig. 4. While the common use of directly movie related words is most apparent in this figure, it can also be seen that Groups 2–5 (which are in the main cluster of Fig. 3) share to various degrees phrases that are either more critical of Marvel/Captain Marvel, involve Alita more, or involve
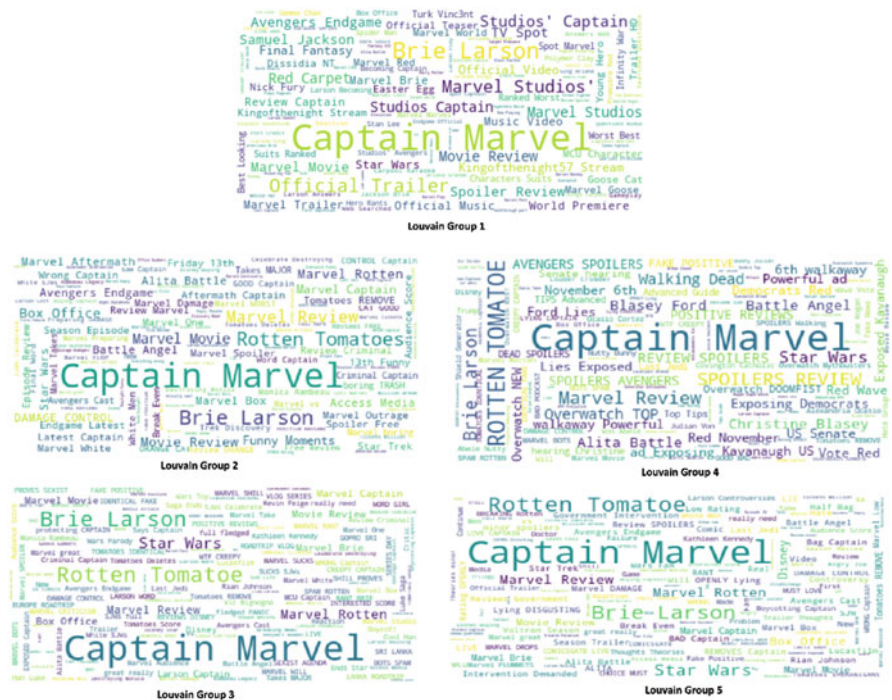
Fig. 4 Word clouds based on YouTube video titles shared by each Louvain group

political/cultural discussion points and targets. This is especially true of Groups 4 and 5 which contain anti-Democrat phrases, pro-Republican phrases (especially in Group 4) and phrases related to other movies with contentious diversity discussions such as Star Wars (Group 4 and 5). This result reinforces our understanding of the main cluster as being where many users that fall on the right-wing/anti-diversity side of such debates are located.

4.3 Types of Actors

We also identified the roles the different tweeters played on each of the conversations. As Table 1 shows the vast majority of actors involved in these conversations were regular Twitter users, followed by news-like agencies and reports and celebrities. Using a conservative bot-score we found low levels of bot-like accounts with the highest level being in the news-like category. Bot-like accounts did not appear to favor one type of interaction over the other (i.e. the percentage of bot-like accounts out of all the accounts that retweeted was similar to those that replied or originated stories), nor did they favor one type of story. Similarly low levels of bot-like account activity was observed in an analysis of a similar event [2].

Table 1 Roles of Twitter users in the four boycott-related campaigns. A bot-score of 0.8 or greater was used to classify bot-like agents

	Only #Ali-taChallenge	Only #BoycottCap-tainMarvel	Both hashtags	Only charged Alita
# Total actors	29,331	6,255	2,167	21,132
By actions (% bot-like):				
Original poster (4%)	3%	7%	10%	21%
Retweeter of original (4%)	33%	61%	51%	40%
Replier of original (5%)	27%	21%	18%	27%
Retweeter of replier (3%)	37%	11%	22%	12%
By identity (% bot-like):				
Regular (4%)	95.5%	96.1%	97.4%	93.4%
News/Reporters (7%)	2.5%	2.3%	1.0%	2.7%
Celebrity (2%)	1.1%	0.8%	1.3%	1.7%
Company (3%)	0.6%	0.5%	0.2%	2.0%
Government (3%)	0.3%	0.3%	0.1%	0.1%
Sports (3%)	0.1%	0.0%	0.0%	0.1%

It should be noted that the #AlitaChallenge discussion while involving the most users, was originated by a smaller percentage of those users when compared to other campaigns. The #BoycottCaptainMarvel campaign had the highest level of actors who were involved in retweeting and spreading the origin tweets.

5   Discussion and Conclusions

We used the interesting case of the boycott-based-on-misinformation campaigns in the Captain Marvel twitter discussion to examine the origins, actors, and diffusion of misinformation. Overall, we found that the origins of all four types of campaigns were located in the same communities that share right-wing/anti-diversity sentiments through YouTube. We found that though the origins and discussion take place in similar communities and with a similar presence of celebrity and news-like accounts, the diffusion of the more indirect #AlitaChallenge and the more direct #BoycottCaptainMarvel occurred in different ways. Here we want to touch upon some of these differences, discuss why they may have come about, and where this research can lead.

#AlitaChallenge and related Charged Alita discussions traveled to a much greater extent than #BoycottCaptainMarvel. This may have occurred because the #AlitaChallenge campaign on the surface presented a negative activity (boycotting Captain Marvel based on misinformation) as a positive one (supporting the strong female lead in Alita). Positive framing of actions has been found in past research to motivate participation, though not always. Our results may more strongly support the idea that #AlitaChallenge was more successful due to (1) successful hijacking

of the hashtag by someone connected to the right-wing/anti-diversity comic book movie community and (2) the coordinated retweeting of a group of similar attempts to push #AlitaChallenge. It should be noted that while the #AlitaChallenge and related campaigns spread more widely than #BoycottCaptainMarvel, they also garnered a much larger response that appears predominately negative. The success of #AlitaChallenge and origin from a specific set of known actors may have been a contributor in making pro-Captain Marvel users more aware and galvanized to respond. The lower level of the #BoycottCaptainMarvel campaign also aligned with a lower level of responses (again most of the users in this discussion were just retweeting the original false-message based campaign).

Our other contribution in this work is demonstrating the possible utility of using both topic groups and cross-platform connections to better define communities of interest. Examining the commonalities in YouTube authors among Twitter users in our data provided a clearer picture of the communities that started and spread misinformation related to the Captain Marvel lead actress. Further examination of the connections between shared YouTube content and Twitter discussions could be useful in looking at the other misinformation being spread about the movie (both from the pro-movie and anti-movie sides). Another future avenue of exploration is to focus on the role those users who cross Twitter topic group or YouTube author clusters play in spreading or responding to specific campaigns.

Additional limitations to this work that could be improved upon in the future include the lack of a more fine-grained picture of the responses to the boycott campaigns, the need for additional Twitter network to YouTube network comparisons, and deeper examination of the cultural context in which these campaigns were taking place.

The purpose of this work should not be taken to be to help in the design of successful misinformation but rather to assist in the understanding of different methods intentionally used to promote false information. This may help with the design of effective and efficient community level interventions – at a minimum as an example of the kinds of campaigns to be aware of and vigilant against. Future work should delve more deeply into the cross-platform and cross-network nature of these conversations with an eye toward how that may improve our ability to classify the intent and effect of various campaigns.

References

1. Amati, G., Angelini, S., Gambosi, G., Pasquin, D., Rossi, G., Vocca, P.: Twitter: temporal events analysis. In: Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good. pp. 298–303. ACM (2018)
2. Babcock, M., Beskow, D.M., Carley, K.M.: Different faces of false: the spread and curtailment of false information in the black panther twitter discussion. J. Data Inf. Qual. (JDIQ) **11**(4), 18 (2019)
3. Benigni, M.C., Joseph, K., Carley, K.M.: Online extremism and the communities that sustain it: detecting the isis supporting community on twitter. PLoS One **12**(12), e0181405 (2017)

4. Bennett, W.L., Livingston, S.: The disinformation order: disruptive communication and the decline of democratic institutions. Eur. J. Commun. **33**(2), 122–139 (2018)
5. Beskow, D.M., Carley, K.M.: Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In: Conference paper. SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (2018)
6. Beskow, D.M., Carley, K.M.: Social cybersecurity: an emerging national security requirement. Mil. Rev. **99**(2), 117 (2019)
7. Bhatt, S., Wanchisen, B., Cooke, N., Carley, K., Medina, C.: Social and behavioral sciences research agenda for advancing intelligence analysis. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 63, pp. 625–627. SAGE Publications: Los Angeles (2019)
8. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: conversational aspects of retweeting on twitter. In: 2010 43rd Hawaii International Conference on System Sciences, pp. 1–10. IEEE (2010)
9. Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A., Van Bavel, J.J.: Emotion shapes the diffusion of moralized content in social networks. Proc. Natl. Acad. Sci. **114**(28), 7313–7318 (2017)
10. Carley, K.M.: Ora: a toolkit for dynamic network analysis and visualization. In: Encyclopedia of Social Network Analysis and Mining, pp. 1219–1228. Springer, New York (2014)
11. Carley, K.M., Cervone, G., Agarwal, N., Liu, H.: Social cyber-security. In: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, pp. 389–394. Springer (2018)
12. Ferrara, E.: Manipulation and abuse on social media by Emilio Ferrara with Ching-man Au Yeung as coordinator. ACM SIGWEB Newsletter (Spring) **4**, 1–9 (2015)
13. Ferrara, E.: Measuring social spam and the effect of bots on information diffusion in social media. In: Complex Spreading Phenomena in Social Systems, pp. 229–255. Springer, Cham (2018)
14. Huang, B.: Learning User Latent Attributes on Social Media. Ph.D. thesis, Carnegie Mellon University (2019)
15. Hussain, M.N., Tokdemir, S., Agarwal, N., Al-Khateeb, S.: Analyzing disinformation and crowd manipulation tactics on youtube. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1092–1095. IEEE (2018)
16. Macskassy, S.A., Michelson, M.: Why do people retweet? anti-homophily wins the day! In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
17. Marwick, A., Lewis, R.: Media Manipulation and Disinformation Online. Data & Society Research Institute, New York (2017)
18. Mejias, U.A., Vokuev, N.E.: Disinformation and the media: the case of Russia and Ukraine. Media Cult. Soc. **39**(7), 1027–1042 (2017)
19. Mueller, R.S.: The Mueller Report: Report on the Investigation into Russian Interference in the 2016 Presidential Election. WSBLD (2019)
20. Proctor, W., Kies, B.: On toxic fan practices and the new culture wars. Participations **15**(1), 127–142 (2018)
21. Ribeiro, M.H., Calais, P.H., Almeida, V.A., Meira Jr, W.: "everything i disagree with is# fakenews": correlating political polarization and spread of misinformation. arXiv preprint arXiv:1706.05924 (2017)
22. Scharl, A., Weichselbraun, A., Liu, W.: Tracking and modelling information diffusion across interactive online media. Int. J. Metadata Semant. Ontol. **2**(2), 135–145 (2007)
23. Snider, M.: Florida school shooting hoax: doctored tweets and Russian bots spread false news. USA Today (2018)
24. Stieglitz, S., Dang-Xuan, L.: Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. J. Manag. Inf. Syst. **29**(4), 217–248 (2013)
25. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: 2010 IEEE Second International Conference on Social Computing, pp. 177–184. IEEE (2010)

26. Tucker, J.A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., Nyhan, B.: Social media, political polarization, and political disinformation: a review of the scientific literature. Political Polarization, and Political Disinformation: A Review of the Scientific Literature (March 19, 2018)
27. Uyheng, J., Carley, K.M.: Characterizing bot networks on twitter: an empirical analysis of contentious issues in the Asia-Pacific. In: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, pp. 153–162. Springer (2019)