There are N Impostors Among Us: Understanding the Effect of State-Sponsored Troll Accounts on Reddit Discussions

Mohammad Hammas Saeed,[1] **Jeremy Blackburn,** [2] **Gianluca Stringhini** [1]

[1] Boston University
[2] Binghamton University
hammas@bu.edu, jblackbu@binghamton.edu, gian@bu.edu

Abstract

Social media platforms are increasingly becoming targets of state-sponsored influence campaigns, carried out through inauthentic accounts known as *troll accounts*. The goal of these campaigns is often to polarize and steer online discussion towards certain strategic narratives. There has been little work, however, on understanding the effect that these influence campaigns have on online discourse, although this is key to assessing their effectiveness and devising efficient countermeasures. In this paper, we study the effect that troll accounts have on online discussions on Reddit. We look at whether these accounts are successful in generating polarized discussions, by comparing the toxicity of comments in threads started by troll accounts compared to general Reddit threads. Our results show that state-sponsored troll accounts on Reddit produce threads that attract more toxic comments than other posts on the same subreddit.

Introduction

In recent years, online polarization and toxicity (Mondal, Silva, and Benevenuto 2017a) as well as disinformation (Mueller 2019; Council 2021) is on the rise on social media platforms. Past research has shown that coordinated campaigns or "raids" are organized by polarized communities such as 4chan's Politically Incorrect Board (/pol/) and "The Donald" (a subreddit on Reddit) to conduct hate speech attacks on different targets (Flores-Saviaga, Keegan, and Savage 2018; Hine et al. 2017; Kumar et al. 2018; Mariconti et al. 2019). A growing number of studies also show that false narratives on social media platforms are an increasingly common problem (Starbird, Arif, and Wilson 2019; Starbird 2017; Vosoughi, Roy, and Aral 2018; Wang et al. 2021). Such activities are often carried out by special accounts controlled by state-sponsored actors called *troll accounts*. Troll accounts on social media interact with one another and appear innocuous to a regular user while covertly being used to spread toxic content and/or disinformation. With increased awareness, researchers have looked into ways these accounts operate and the agendas they push, for example, by studying the troll accounts identified by Reddit and Twitter and active from 2014 to 2018 (Bessi and Ferrara 2016; Ferrara 2017; Xia et al. 2019; Zannettou et al. 2019a,b,c).

While researchers have investigated hate speech and toxic discourse enacted by troll accounts on social media (Chatzakou et al. 2017; Davidson et al. 2017; ElSherief et al. 2018b), there is little work done on understanding the success of their actions and the impact they are creating. More specifically, it is not documented how the conversations of troll accounts differ from those of regular users, i.e., whether troll accounts foster conversations that are more toxic than the regular ones on the platform. Therefore, to bridge this gap, we aim to answer the following research question in this paper:

- **RQ – *Thread Toxicity:*** Are comments in threads started by trolls more toxic than comments in general threads?

To answer the research question, we use data from 335 Russian-sponsored troll accounts released by Reddit. We collect all the posts made by these accounts and the comments made on those posts. We establish a baseline of threads in each of the Top-15 subreddits where trolls post in order to compare the toxicity with troll threads.

Our results show that for each of the Top-15 subreddits that trolls post in, the overall toxicity is higher for troll threads than regular threads. We validate our results by performing a z-test on toxicity scores for each subreddit and find that the difference in toxicity is statistically significant for all subreddits.

Disclaimer. The goal of troll accounts is to generate controversy and polarize online discussion. As such, the language of the posts that we analyze in this paper often contains profanity and slurs. In the examples in this paper, we do not censor any language, so we warn the reader that they might find some of this content upsetting.

Related Work

In this section, we discuss previous work on troll accounts, their reach on social media platforms, and hate speech.

Hate Speech. Silva et al. (Silva et al. 2016) examine the content shared on Twitter and Whisper to identify the primary targets of hate speech on social media platforms. Mondal et al. (Mondal, Silva, and Benevenuto 2017b) investigate common hate expressions and the impact of anonymity on hate speech. They also provide information on the most

hated groups on the internet. Mathew et al. (Mathew et al. 2019) examine Gab content for hate speech and discover that the most toxic content spreads the fastest and farthest. ElSherief et al. (ElSherief et al. 2018b,a) investigate the targets of hate speech on Twitter as well as the perpetrators of toxic content. Finkelstein et al. (Finkelstein et al. 2018) discuss the growing use of racial slurs on 4chan and Gab, with a focus on anti-semitism. Chandrasekharan et al. (Chandrasekharan et al. 2017) investigate the shift in hate speech following the banning of prominent toxic subreddits such as r/fatpeople and r/CoonTown. They discovered that banning these subreddits resulted in a decrease in hate speech because accounts that posted toxic content either migrated to another platform or stopped posting entirely. Olteanu et al. (Olteanu et al. 2018) investigate hate speech on the internet caused by real-world extremist attacks by Arabs and Muslims, concluding that hate speech (particularly violent content) increases after such incidents. Jhaver et al. (Jhaver et al. 2018) investigate the effect of blocklists on online harassment. They discover that users either believe they have been unfairly blocked or that they are unprotected online. Erjavec and Kovacic (Erjavec and Kovačič 2012) in an attempt to understand the motivations and strategies of hate speech posters through interviews. Some hate speech posters are part of organized campaigns, whereas others are frequently motivated by thrill and fun. Hughey and Daniels (Hughey and Daniels 2013) investigate the strategies used by news platforms to study racist comments. They examine the drawbacks and implications of methods such as extreme moderation policies, comment disabling, and so on. Harlow (Harlow 2015) study of comments on US news sites (e.g. racial slurs) to better understand racist discourse. They discover that Latinos are the most targeted ethnicity, and racial slurs are mentioned in comments even when the article contains none. Zollo et al. (Zollo et al. 2015) examine Facebook data and discover that discussions about conspiracy theories are more negative than those about science. Finally, Zannettou et al. (Zannettou et al. 2018) investigate the spread of hateful memes on the Internet.

Troll Activity on Social Media. Zannettou et al. (Zannettou et al. 2019b,c) investigate state-sponsored troll accounts active on Twitter and Reddit between 2014 and 2018. They investigate how successfully these accounts were able to spread their content on those platforms as well as other Web communities and discover that troll accounts are typically created in waves. The same authors also created an analysis pipeline to study images posted by troll accounts on Twitter (Zannettou et al. 2019a). Volkova and Bell (Volkova and Bell 2016) look at 180k Twitter accounts that were active during the Russia-Ukraine conflict. They discover that lexical features are strong predictors of whether a Twitter account will be flagged as a troll and suspended as a result. Luceri et al. (Luceri, Giordano, and Ferrara 2020) uses Inverse Reinforcement Learning to detect troll accounts on Twitter (IRL). Bot detection systems have previously used the same features used by Luceri et al. to detect trolls. Kumar et al. (Kumar et al. 2017) investigate attempts to manipulate users' opinions on social media platforms using a set of accounts known as sockpuppets that are controlled by the same user. According to Mihaylov and Nakov (Mihaylov and Nakov 2016), there are two types of troll accounts: 1) paid accounts used to spread a specific message, and 2) accounts that act on their own volition. Mihaylov et al. (Mihaylov, Georgiev, and Nakov 2015) later demonstrate that trolls do indeed manipulate users' opinions on online forums. Steward et al. (Steward, Arif, and Starbird 2018) analyze the activity of Russian-sponsored trolls on Twitter during the Black Lives Matter debate. They discover that these accounts pushed specific narratives on both left and right-leaning communities. Varol et al. (Varol et al. 2017) propose a system for labeling memes that became popular as a result of collaborative efforts. Using machine learning techniques, Ratkiewicz et al. (Ratkiewicz et al. 2011) detect the dissemination of false political information on Twitter. Howard and Kollanyi (Howard and Kollanyi 2016) investigate the bots that were active during the 2016 Brexit referendum and discover that they primarily promoted pro-Brexit narratives. They also reveal that 1% of the accounts posted 33% of the messages. Hegelich and Janetzko (Hegelich and Janetzko 2016) analyze 1.7k Twitter bots active during the Russia-Ukraine conflict. They reveal the political agendas of these bots as well as the behaviors associated with these accounts, such as hiding one's identity, using hashtags to push narratives, and retweeting specific content. Badawy et al. (Badawy, Lerman, and Ferrara 2018) conduct research on state-sponsored actors and predict whether they will spread misinformation. Dutt et al. (Dutt, Deb, and Ferrara 2018) investigate Facebook ads shared by Russian troll accounts and the characteristics that make such strategies effective.

Data

For the purposes of this study, we use data released by Reddit on Russian-sponsored troll accounts active between 2015 and 2018 (Reddit 2017). This dataset comprises 335 accounts, which made a total number of 14,224 posts, which attracted 88,502 comments from other Reddit users. It is important to note that these troll accounts were active during some of the major political events, such as the 2016 Brexit Referendum, the 2016 US Presidential Election, and the 2018 US Midterm Election, which makes the dataset more interesting to study. To conduct our analysis, we need a baseline of threads to compare their toxicity with that of troll threads. Therefore, we download Pushshift's public archives (Baumgartner et al. 2020) which includes all public posts and comments made on Reddit from 2005 to 2020. This archive contains 600M posts and 5B comments on 2.8M subreddits (Baumgartner et al. 2020).

Reddit. Reddit is one of the most popular sites for news discussion (Samory and Mitra 2018; Weninger, Zhu, and Han 2013; Weninger 2014; Zannettou et al. 2017). More broadly, Reddit is characterized as a social news aggregation, content rating, and discussion website. On Reddit, content is organized into communities made by users called subreddits, where each subreddit is targeted towards a certain topic (e.g. news, jokes etc). In a subreddit, a user can create a thread called a submission and other users can reply to it by posting comments. Users can reply to the original submission or

to another user's comment.

Ethics. Our work is restricted to publicly available data and there is no interaction with human subjects, therefore it is not considered human subjects research by the IRB at our institution. Also, going by the standard ethics guidelines, we do not further deanonymize users and remove any PII in the examples that we provide.

Analysis

In this section, we aim to answer our research question. To conduct our analyses, we have to calculate the toxicity of Reddit comments. For that, we use Google's Perspective API, a free Google service developed by Jigsaw (Google 2020). The Perspective API uses a machine learning model trained on comments manually labeled as toxic or non-toxic (Delgado 2019). The API returns several scores, including "Toxicity" and "Severe Toxicity" where each score ranges from 0 to 1. We use "Severe Toxicity" as our metric, since prior work (Zannettou et al. 2020; Jhaver et al. 2021; Fortuna, Soler, and Wanner 2020) shows it to be a more robust indicator of hateful speech. For brevity, "Severe Toxicity" is referred to as "toxicity" throughout the section.

RQ: Are comments in threads started by trolls more toxic than general thread comments?

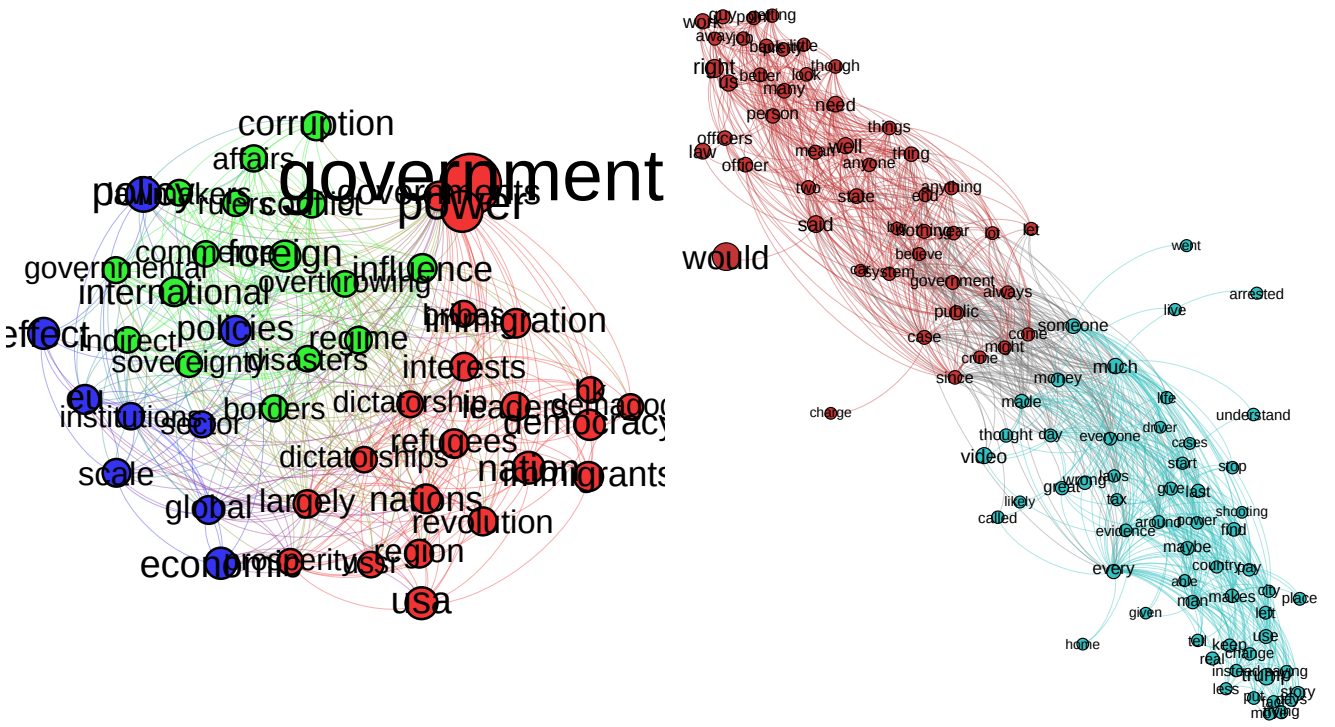
The main purpose of this research question is to determine whether troll accounts are able to create Reddit threads that attract more toxic comments than regular threads. This will give us a clearer understanding of the impact these troll accounts have on the subreddits they post on and the kind of atmosphere they create. We first extract the Top-15 subreddits on which troll accounts post. We select the Top-15 subreddits because this gives us enough data to perform relevant analysis while also adhering to the quota limitations imposed on us by the Perspective API. Table 1 shows all subreddits, which range from r/blackpower to r/politics. Many subreddits are related to politics, news, and race, which makes them ideal grounds for polarizing discussions. We then use Google's Perspective API to calculate the toxicity of all comments made in the threads started by trolls in those subreddits. Next, we calculate the average toxicity of troll threads in a given subreddit. To establish a baseline toxicity for each subreddit, we select a random set of the same number of threads as the trolls. We also ensure that the comments are made during the same time period so that the only differentiating factor is the toxicity. Table 1 shows the results of our experiment, and for all subreddits, it is evident that troll threads are much more toxic than the baseline toxicity. We also perform a z-test for each subreddit to show that troll threads are more toxic than general threads. To calculate the z-score, we use the mean toxicity from each row of Table 1 as proportion, where the population size is the number of threads in the subreddit. Our results show that for all the subreddits, the differences in toxicity show statistically significant differences ($p < 0.01$), allowing us to answer RQ in the affirmative.

Language Usage. In this section, we compare the use of language between troll threads and the general threads in the same subreddits. We highlight the difference on three important keywords "black," "government," and "trump." We select these keywords for two reasons: 1) these keywords are the basis of many polarizing discussions by trolls, and 2) these keywords appear in the Top-100 words in troll threads. The word "black" appears in 1,950 comments under troll posts, "government" appears in 1,429 comments, and "trump" appears in 1,165 comments. We use the methodology by Zannettou et al. (Zannettou et al. 2019d) to visualize the language in relation to each keyword. Figure 1 present the graphs calculated from the word "government", Figure 2 present the graphs calculated from the word "black" and Figure 3 present the graphs calculated from the word "trump." Each word is a node and is connected by an edge if the cosine similarity of their embedding vectors is above a given threshold. The threshold for trolls is set to 0.9 for "black", "trump" and "government," whereas for the general comments, the threshold is set to 0.79, 0.87 and 0.84 respectively. These thresholds are selected to keep approximately 50-100 nodes in each graph. We chose this range to have a reasonable number of nodes for visualization.

We perform a series of steps to visualize the graphs. First, we create a weighted graph with the ForceAtlas2 layout algorithm (Jacomy et al. 2014), in which words with higher cosine similarities are arranged closer together in the graph space. We also run the Louvain community detection algorithm (Blondel et al. 2008) on the graph to identify "communities" of similar words as used by past research (Papasavva et al. 2021). Words from the same community are represented by the same color. Figure 1a shows the word embedding graph for the keyword "government" in troll threads and Figure 1b those of general threads. As it can be seen, troll threads have discussions related to far more polarizing topics such as "dictatorship, overthrowing, bribes, corruption" whereas the general threads have much more benign topics such as "tax, money, law." Similarly, Figure 2a shows the word embedding graph for the keyword "black"

Subreddit	Troll Threads Toxicity	Baseline Toxicity	Z-Score	P-Value
Bad_Cop_No_Donut	0.22	0.04	10.43	<.00001
The_Donald	0.22	0.04	6.38	<.00001
blackpower	0.22	0.04	5.37	<.00001
news	0.21	0.05	5.92	<.00001
Blackfellas	0.21	0.04	4.11	<.00001
POLITIC	0.19	0.03	5.97	<.00001
copwatch	0.18	0.03	4.84	<.00001
interestingasfuck	0.17	0.03	4.26	<.00001
police	0.17	0.03	4.16	<.00001
gifs	0.16	0.03	5.70	<.00001
uspolitics	0.16	0.03	4.02	<.00001
racism	0.16	0.03	5.27	<.00001
Health	0.15	0.02	3.76	0.00016
PoliticalHumor	0.14	0.02	7.07	<.00001
politics	0.13	0.01	4.42	<.00001

Table 1: For each of the top subreddits, we compare the baseline toxicity with the toxicity of troll threads.



(a) Troll threads

(b) General threads

Figure 1: The graph depicts language usage for the keyword "government" with nodes from the same community depicted in the same color and detected using the Louvain community detection method (Blondel et al. 2008). The narratives in trolls threads are clearly more extreme than in general threads.

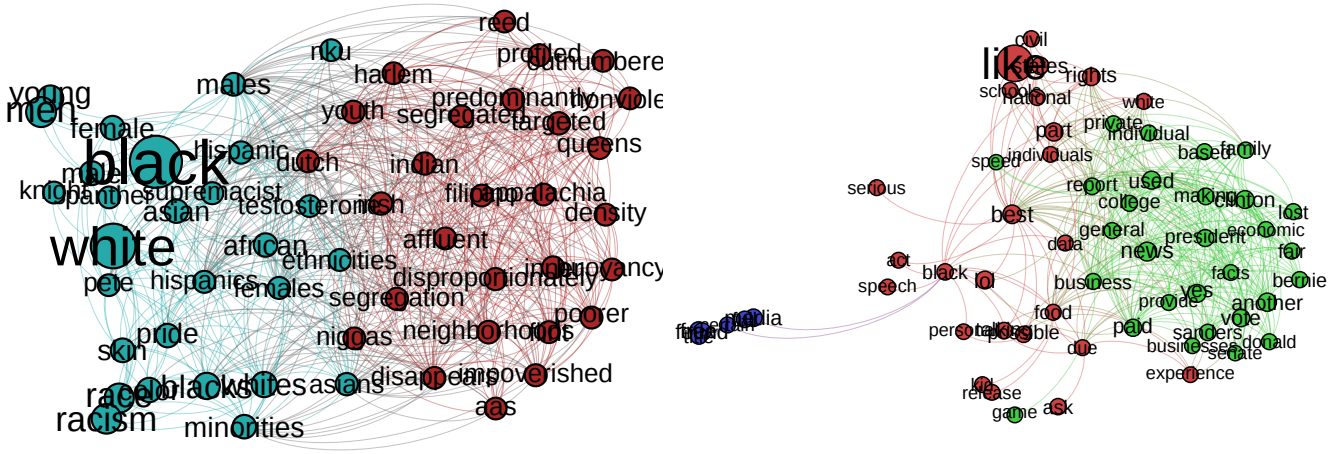
in troll threads and Figure 2b those of general threads. Troll threads focus on topics such as "supremacy, dehumanizing, racism, criminality, injustice, abuse, thug, enslavement." On the other hand, we don't see the same extremism in the general discussion of the keyword and there is some mention of "rights" and "speech." Figure 3a shows the word embedding graph for the keyword "trump" in troll threads and Figure 3b those of general threads. Troll threads have more polarizing topics occurring e.g. "rigged, cruel, scandals" and general threads have more general topics such as "system, country, government."

Comment Examples. To further illustrate the toxicity of language and topics covered by trolls, we discuss a few, manually selected, comments containing the word "black." These examples are taken from comments posted by regular users on troll threads to show that troll threads can host highly toxic conversations.

> COMMENT 1: GAW DON'T YOU GET IT BLACK PEOPLE ARE TERRIFIED OF COPS CAUSE THEY ALL RACIST N SHIT, MAYNE FUKIN PIGS MAYNE
> COMMENT 2: BLM = black racists who genuinely believe their own bullshit
> COMMENT 3: i think they're actually kind of funny. and they do accurately mock a portion of the black community. i work in a warehouse and there are few entitled bitchy black chicks that are such a fucking pain in the ass. you cant even look in their general direction without them trying to start shit.

All three comments clearly show a racist narrative and the extreme use of language. Many comments with high toxicity cover topics such as racism which was also evident previously in Figure 2a.

Takeaways. Overall, our analysis shows that comments in trolls threads are more toxic than general thread comments. We find that for all of the top subreddits, troll threads are much more toxic and derive discussions that lead to racism,



(a) Troll threads

(b) General threads

Figure 2: The graph depicts language usage for the keyword "black" with nodes from the same community depicted in the same color and detected using the Louvain community detection method (Blondel et al. 2008). The narratives in trolls threads are clearly more extreme than in general threads.

political extremism, and hate speech.

Discussion and Conclusion

In this paper, we seek an answer to one major research question: 1) whether troll accounts generate more toxic conversations than regular accounts. Our results show that on average, the threads started by trolls cause much more toxic conversations as compared to general threads.

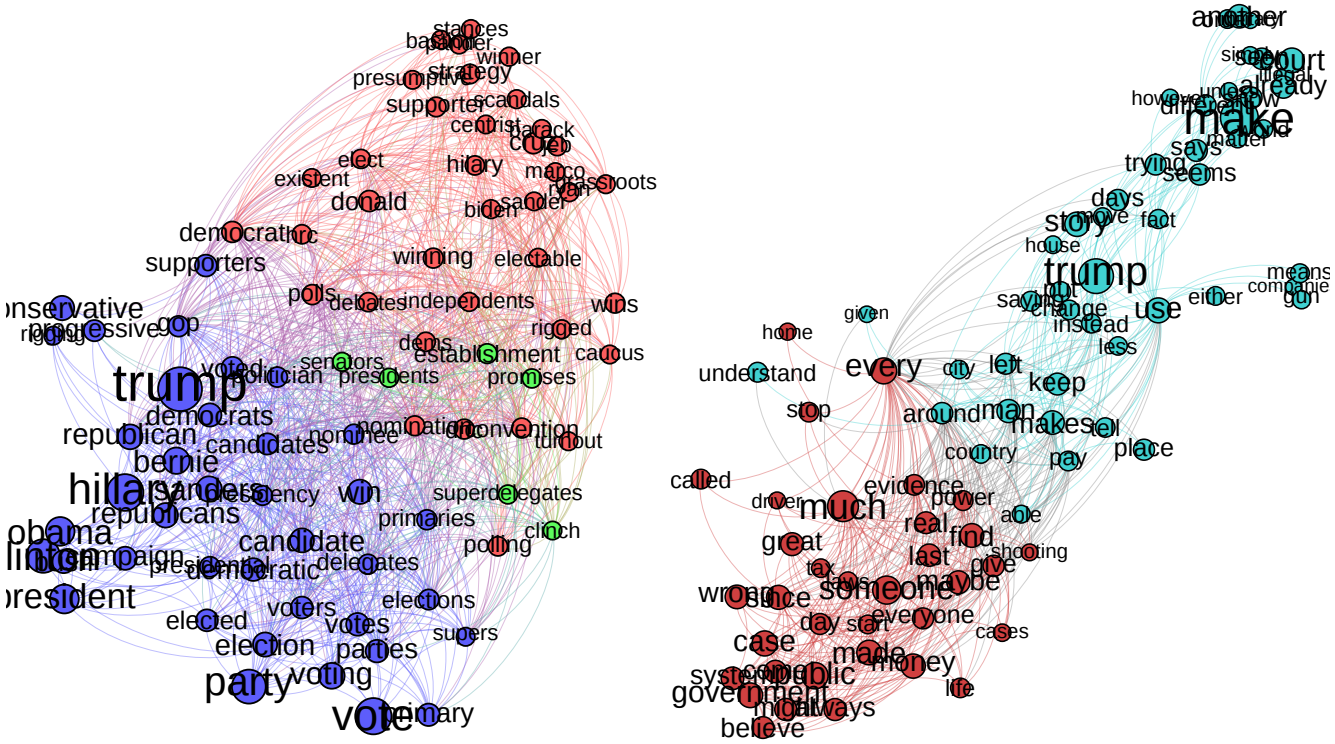
Implications for computational social science research and potential future work. Our work has interesting implications for future research in understanding the effect of state-sponsored troll operations on online user polarization. In particular, we find that threads started by troll accounts attract more toxic discussion. We believe that an interesting line of research that should be pursued is analyzing the online activity of trolls and users who interact with them qualitatively, as well as conducting interview studies with people who interacted with trolls to understand how their attitudes might have shifted.

Implications for social platforms. Our work shows that state-sponsored troll accounts are successful in generating polarized online discussions that attract considerable toxic speech. This indicates that, unlike other types of malicious content like spam, which is mostly ignored by users and does not attract interactions, the content posted by trolls has the potential to cause considerable harm the longer it is left on the online platform. For this reason, developing effective systems to detect and take down inauthentic content posted by state-sponsored actors is of paramount importance for having safe and thriving online communities.

Acknowledgments. This work was supported by the NSF under grants CNS-1942610, IIS-2046590, CNS-2114407, and CNS-2114411.

References

Badawy, A.; Lerman, K.; and Ferrara, E. 2018. Who Falls for Online Political Manipulation? *arXiv preprint arXiv:1808.03281*.



(a) Troll threads

(b) General threads

Figure 3: The graph depicts language usage for the keyword "trump" with nodes from the same community depicted in the same color and detected using the Louvain community detection method (Blondel et al. 2008). The narratives in trolls threads are clearly more extreme than in general threads.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 830–839.

Bessi, A., and Ferrara, E. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21(11).

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW).

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017. Mean birds: Detecting aggression and bullying on twitter. In *ACM Conference on Hypertext and Social Media*.

Council, N. I. 2021. *Foreign Threats to the 2021 US Federal Elections*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *AAAI International Conference on Web and Social Media (ICWSM)*.

Delgado, P. 2019. How El País used AI to make their comments section less toxic. https://blog.google/outreach-initiatives/google-news-initiative/how-el-pais-used-ai-make-their-comments-section-less-toxic/.

Dutt, R.; Deb, A.; and Ferrara, E. 2018. 'Senator, We Sell Ads': Analysis of the 2016 Russian Facebook Ads Campaign. *arXiv preprint arXiv:1809.10158*.

ElSherief, M.; Kulkarni, V.; Nguyen, D.; Wang, W. Y.; and Belding, E. M. 2018a. Hate lingo: A target-based linguistic analysis of hate speech in social media. *CoRR* abs/1804.04257.

ElSherief, M.; Nilizadeh, S.; Nguyen, D.; Vigna, G.; and Belding, E. M. 2018b. Peer to peer hate: Hate speech instigators and their targets. *CoRR* abs/1804.04649.

Erjavec, K., and Kovačič, M. P. 2012. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society* 15(6):899–920.

Ferrara, E. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *ArXiv 1707.00086*.

Finkelstein, J.; Zannettou, S.; Bradlyn, B.; and Blackburn, J. 2018. A quantitative approach to understanding online antisemitism. *CoRR* abs/1809.01644.

Flores-Saviaga, C.; Keegan, B. C.; and Savage, S. 2018. Mobilizing the trump train: Understanding collective action in a political trolling community. In *AAAI International Conference on Web and Social Media (ICWSM)*.

Fortuna, P.; Soler, J.; and Wanner, L. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 6786–6794. Marseille, France: European Language Resources Association.

Google. 2020. Perspective API. https://www.perspectiveapi.com.

Harlow, S. 2015. Story-chatterers stirring up hate: Racist discourse in reader comments on u.s. newspaper websites. *Howard Journal of Communications* 26(1):21–42.

Hegelich, S., and Janetzko, D. 2016. Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet. In *AAAI International Conference on Web and Social Media (ICWSM)*.

Hine, G. E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *AAAI International Conference on Web and Social Media (ICWSM)*.

Howard, P. N., and Kollanyi, B. 2016. Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. *CoRR* abs/1606.06356.

Hughey, M. W., and Daniels, J. 2013. Racist comments at online news sites: a methodological dilemma for discourse analysis. *Media, Culture & Society* 35(3):332–347.

Jacomy, M.; Venturini, T.; Heymann, S.; and Bastian, M. 2014. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOS ONE* 9(6):1–12.

Jhaver, S.; Ghoshal, S.; Bruckman, A.; and Gilbert, E. 2018. Online harassment and content moderation: The case of blocklists. *ACM Trans. Comput.-Hum. Interact.* 25(2).

Jhaver, S.; Boylston, C.; Yang, D.; and Bruckman, A. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. *Proc. ACM Hum.-Comput. Interact.* 5(CSCW2).

Kumar, S.; Cheng, J.; Leskovec, J.; and Subrahmanian, V. 2017. An army of me: Sockpuppets in online discussion communities. In *The Web Conference (WWW)*.

Kumar, S.; Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, 933–943. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

Luceri, L.; Giordano, S.; and Ferrara, E. 2020. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *ICWSM*.

Mariconti, E.; Suarez-Tangil, G.; Blackburn, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Serrano, J. L.; and Stringhini, G. 2019. "you know what to do": Proactive detection of youtube videos targeted by coordinated hate attacks. In *Proceedings of the ACM on Human Computer Interaction (CSCW)*.

Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, 173–182. New York, NY, USA: Association for Computing Machinery.

Mihaylov, T., and Nakov, P. 2016. Hunting for Troll Comments in News Community Forums. In *ACL*.

Mihaylov, T.; Georgiev, G.; and Nakov, P. 2015. Finding Opinion Manipulation Trolls in News Community Forums. In *CoNLL*.

Mondal, M.; Silva, L. A.; and Benevenuto, F. 2017a. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, 85–94. New York, NY, USA: Association for Computing Machinery.

Mondal, M.; Silva, L. A.; and Benevenuto, F. 2017b. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, 85–94. New York, NY, USA: Association for Computing Machinery.

Mueller, R. S. 2019. *The Mueller report: Report on the investigation into Russian interference in the 2016 presidential election*. WSBLD.

Olteanu, A.; Castillo, C.; Boy, J.; and Varshney, K. R. 2018. The effect of extremist violence on hateful speech online. *CoRR* abs/1804.05704.

Papasavva, A.; Blackburn, J.; Stringhini, G.; Zannettou, S.; and De Cristofaro, E. 2021. "is it a qoincidence?": An exploratory study of qanon on voat. 460–471.

Ratkiewicz, J.; Conover, M.; Meiss, M. R.; Gonçalves, B.; Flammini, A.; and Menczer, F. 2011. Detecting and Tracking Political Abuse in Social Media. In *AAAI International Conference on Web and Social Media (ICWSM)*.

Reddit. 2017. Reddit's 2017 transparency report and suspect account findings. https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/.

Samory, M., and Mitra, T. 2018. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *AAAI International Conference on Web and Social Media (ICWSM)*.

Silva, L. A.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2016. Analyzing the targets of hate in online social media. *CoRR* abs/1603.07709.

Starbird, K.; Arif, A.; and Wilson, T. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*.

Starbird, K. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *AAAI International Conference on Web and Social Media (ICWSM)*.

Steward, L.; Arif, A.; and Starbird, K. 2018. Examining Trolls and Polarization with a Retweet Network. In *MIS2*.

Varol, O.; Ferrara, E.; Menczer, F.; and Flammini, A. 2017. Early detection of promoted campaigns on social media. *EPJ Data Science*.

Volkova, S., and Bell, E. 2016. Account Deletion Prediction on RuNet: A Case Study of Suspicious Twitter Accounts Active During the Russian-Ukrainian Crisis. In *NAACL-HLT*.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*.

Wang, Y.; Tamahsbi, F.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Magerman, D.; Zannettou, S.; and Stringhini, G. 2021. Understanding the use of fauxtography on social media. In *AAAI International Conference on Web and Social Media (ICWSM)*.

Weninger, T.; Zhu, X. A.; and Han, J. 2013. An exploration of discussion threads in social news sites: A case study of the reddit community. In *ASONAM*.

Weninger, T. 2014. An exploration of submissions and discussions in social news: Mining collective intelligence of reddit. *Social Network Analysis and Mining*.

Xia, Y.; Lukito, J.; Zhang, Y.; Wells, C.; Kim, S. J.; and Tong, C. 2019. Disinformation, performed: Self-presentation of a russian ira account on twitter. *Information, Communication and Society*.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *ACM SIGCOMM Internet Measurement Conference (IMC)*.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *ACM SIGCOMM Internet Measurement Conference (IMC)*.

Zannettou, S.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2019a. Characterizing the use of images by state-sponsored troll accounts on twitter. In *AAAI International Conference on Web and Social Media (ICWSM)*.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019b. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *WWW Companion*.

Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019c. Who let the trolls out?: Towards understanding state-sponsored trolls. In *ACM Conference on Web Science*.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2019d. A quantitative approach to understanding online antisemitism.

Zannettou, S.; ElSherief, M.; Belding, E.; Nilizadeh, S.; and Stringhini, G. 2020. Measuring and Characterizing Hate Speech on News Websites. In *ACM WebSci*.

Zollo, F.; Novak, P. K.; Del Vicario, M.; Bessi, A.; Mozetič, I.; Scala, A.; Caldarelli, G.; and Quattrociocchi, W. 2015. Emotional dynamics in the age of misinformation. *PLOS ONE* 10(9):1–22.