AFRL-AFOSR-VA-TR-2021-0067

Digital Deception: The Cognitive and Social Mechanisms of the Spread of Fake News

Lerman, Kristina
UNIVERSITY OF SOUTHERN CALIFORNIA
3720 S FLOWER ST FL 3
LOS ANGELES, CA, 90007
USA

07/07/2021
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
07-07-2021	Final	01 Sep 2019 - 31 Dec 2020

4. TITLE AND SUBTITLE
Digital Deception: The Cognitive and Social Mechanisms of the Spread of Fake News

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-17-1-0327

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Kristina Lerman

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF SOUTHERN CALIFORNIA
3720 S FLOWER ST FL 3
LOS ANGELES, CA 90007
USA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR RTA2

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-VA-TR-2021-0067

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project investigates coupled sensor configuration and planning (CSCP) for autonomous systems, which is a mode of active control of information in context to a decision-making problem. We consider a scenario consisting of a network of mobile vehicles called sensors, and a separate network of mobile vehicles called actors. Sensors gather information about the environment whereas actors perform desired tasks. Specifically, the actors perform tasks encoded in terms of multi-vehicle route-planning problems in a threatening environment. The threat is an unknown spatiotemporally-varying scalar field that is estimated using observations made by sensors. The major accomplishments and successes of this project are as follows:

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			LAURA STECKMAN
U	U	U	UU	7	**19b. TELEPHONE NUMBER** *(Include area code)* 426-7556

FINAL REPORT
Project Title: Digital Deception: The Cognitive and Social Mechanisms of the Spread of Fake News

AFOSR Grant No. FA9550-17-1-0327
Performance Period: September 2017 – March 2021.
Principal Investigator: Dr. Kristina Lerman (lerman@isi.edu)
Lead Institution: University of Southern California Information Sciences Institute
AFOSR Program Manager: Laura Steckman

Summary of Objectives and Outcomes:

The spread of disinformation online highlights our vulnerability to digital influence. However, it remains unclear how to identify online manipulation and mitigate its deleterious effects. The goal of this effort is to elucidate how misinformation spreads online. The research program combines cognitive science with network analysis and data science to create computational models to study influence campaigns using large-scale social media data. The computational models will help characterize, identify and predict the spread of digital deceptions. Our analysis of *the 2016 Russian Internet Research Agency (IRA) disinformation campaign* identified how foreign influence operations targeted social media users based on their political ideology. Our work also elucidated the cognitive factors and linguistic features of the disinformation campaign, which can be used to identify future influence campaigns. Our analysis of *social bots*, i.e., automated accounts widely used to amplify social influence campaigns, helped to more accurately identify online automated accounts. Our work also elucidated the degree to which the *structure of social networks* can distort the perceptions of popularity to social media users, highlighting a new vulnerability in online influence campaigns.

The Covid-19 pandemic provides an important case study for this project. Tragically, many Americans dismissed the pandemic as a hoax and refused to adopt preventive measures, such as social distancing and face covering. The rampant misinformation around the pandemic amplified political polarization and reduced confidence in our institutions, including trust in health experts at the local and national level. To better understand the role of social media misinformation in social polarization, we have collected a large corpus of data from Twitter. The ongoing collection makes available to the research community over a billion messages related to the Covid-19 pandemic. The data is providing rich grounds for analysis of digital influence, including dynamics of polarization and the role of bots in spreading conspiracies. We developed methods to study online polarization, focusing on attitudes toward science. Using geo-referenced social media data, we were able to explore the role of socio-economic factors in shaping skepticism toward science.

Accomplishments and Findings

The following sections summarize key accomplishments and findings in each of our main challenge areas. Results in each challenge area are grouped into one or more topics.

Cognitive Factors in Online Manipulation: the 2016 IRA influence campaign
Key paper: Addawood, A., Badawy, et al. (2019); Badawy, Addawood et al, (2019); Badawy et al (2019)

To study the Internet Research Agency (IRA) manipulation campaign in the 2016 US presidential elections, we collected tweets from accounts associated with the identified *Russian trolls* as well as users sharing posts in the same time period on a variety of topics around the 2016 elections. We used label propagation to infer the users' ideology based on the news sources they share. We were able to

classify a large number of the users as liberal or conservative with precision and recall above 84%. Conservative users who retweet Russian trolls produced significantly more tweets than liberal ones, about 8 times as many in terms of tweets. Additionally, trolls' position in the retweet network is stable overtime, unlike users who retweet them who form the core of the election-related retweet network by the end of 2016. Using state-of-the-art bot detection techniques, we estimated that about 5% and 11% of liberal and conservative users are bots, respectively. Text analysis on the content shared by trolls reveal that conservative trolls talk about refugees, terrorism, and Islam, while liberal trolls talk more about school shootings and the police. Although an ideologically broad swath of Twitter users were exposed to Russian trolls in the period leading up to the 2016 U.S. Presidential election, it is mainly conservatives who help amplify their message.

To study how IRA trolls attempted to manipulate public opinion, we identified 49 theoretically grounded linguistic markers of deception and measured their use by troll and non-troll accounts. We show that deceptive language cues can help to accurately identify trolls, with average F1 score of 82% and recall 88%. In addition, we examine the features of users who play a role in spreading the malicious content created by Russian trolls. We used these features to construct machine learning models that are able to very accurately identify users who spread the trolls' content (average AUC score of 96%, using 10-fold validation). We show that political ideology, bot likelihood scores, and some activity-related account metadata are the most predictive features of whether a user spreads trolls' content or not.

In addition, we examined the features of users who play a role in spreading the malicious content created by Russian trolls. We used these features to construct machine learning models that are able to very accurately identify users who spread the trolls' content (average AUC score of 96%, using 10-fold validation). We show that political ideology, bot likelihood scores, and some activity-related account metadata are the most predictive features of whether a user spreads trolls' content or not.

Bot Activity in Online Discussions:
Key papers: Ferrara (2020), Pozanna & Ferrara (2020), Luceri et al. (2019), Yang et al (2019), Stella, Ferrara & De Domenico (2018), Badawy, Lerman & Ferrara (2019)

We continued to examine the ways in which social bots—automated or semi-automated accounts designed to impersonate humans—have been manipulating online discourse. As bots become more sophisticated, and to some extent capable of emulating the short-term behavior of human users, we must develop methods to automatically identify them. We analyzed the behavioral dynamics that bots exhibit over the course of an activity session to highlight how these differ from human activity. By using a large Twitter dataset associated with recent political events, we were able to separate bots and humans. Our analysis shows the presence of short-term behavioral trends in humans, which can be associated with a cognitive origin, that are absent in bots, intuitively due to the automated nature of their activity. These findings are finally codified to create and evaluate a machine learning algorithm to detect activity sessions produced by bots and humans, to allow for more nuanced bot detection strategies.

In addition, our COVID-19 data set provides early evidence of the use of bots to promote political conspiracies in the United States, in stark contrast with people who focus on public health concerns.

Impact of Social Network Structure on Individual Perceptions
Key papers: Alipourfard et al (2020), Ngo et al. (2020)

We published two important papers, one in *Nature Communications*, and one in the *Journal of Royal Society A*. These papers address the question of how the structure of networks shapes the perceptions of/information received by individual nodes.
Social networks shape perceptions by exposing people to the actions and opinions of their peers. However, the perceived popularity of an opinion may be very different from its actual popularity, especially in online social networks, where people only see their friends' posts. We attribute this perception bias to friendship paradox and identify conditions under which it appears. We validate the findings empirically using Twitter data. Within posts made by users in our sample, we identify topics that appear more often within users' social feeds than they do globally among all posts. We also present a polling algorithm that leverages the friendship paradox to obtain a statistically efficient estimate of a topic's global prevalence from biased individual perceptions. We characterize the polling estimate and validate it through synthetic polling experiments on Twitter data, providing a framework for unbiased (or less biased) estimation of opinions.

Our work has also explicated the role of network structure in perception bias. While past research has shown that some processes on networks may be characterized by local statistics describing nodes and their neighbours, such as degree assortativity, these quantities fail to capture important sources of variation in network structure. We define a property called transsortativity that describes correlations among a node's neighbours (see Fig 1). Transsortativity can be systematically varied, independently of the network's degree distribution and assortativity. Moreover, it can significantly impact the spread of contagions as well as the perceptions of neighbours, known as the majority illusion.

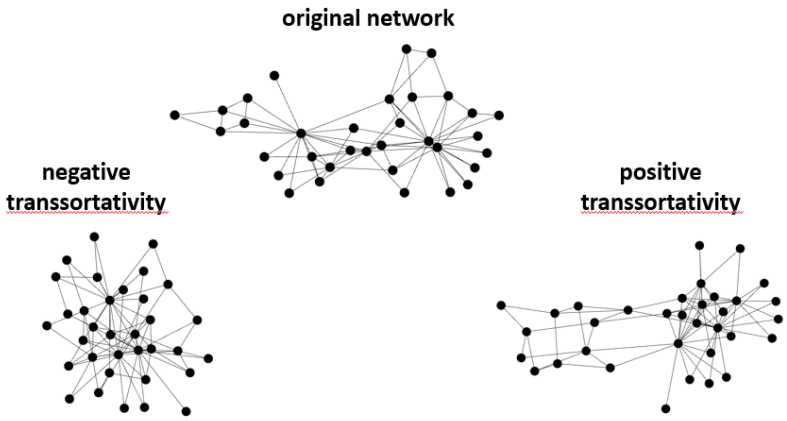

Figure 1: New network property identified by Lerman and collaborators, which they called transsortativity. Transsortativity measures degree correlations of a node's neighbors. The plot on the right shows the karate club social network (center) rewired to have positive transsortativity, while keeping all other network properties (degree distribution, degree assortativity) fixed. The plot on the left shows the same network rewired for negative transsortativity, while keeping all other properties fixed.

Polarization in Online Discussions about Covid-19
Key papers: Chen et al. (2020), Jiang et al. (2020), Rao et al. (2021), Hu et al. (2021)

We started an ongoing data collection of tweets related to the Covid-19 pandemic on January 28, 2020. In the first release, we have published over 123 million tweets, with over 60% of the tweets in English. (As of this writing, there are over one billion tweets in the data set.)

By linking 2.3 million Twitter users tweeting in English to locations within the United States, we were able to show in aggregate that COVID-19 chatter in the United States is largely shaped by political polarization. Partisanship correlates with sentiment toward government measures and the tendency to share health and prevention messaging. Cross-ideological interactions are modulated by user segregation and polarized network structure. We also observe a correlation between user engagement with topics related to public health and the impact of the disease outbreak in different U.S. states.

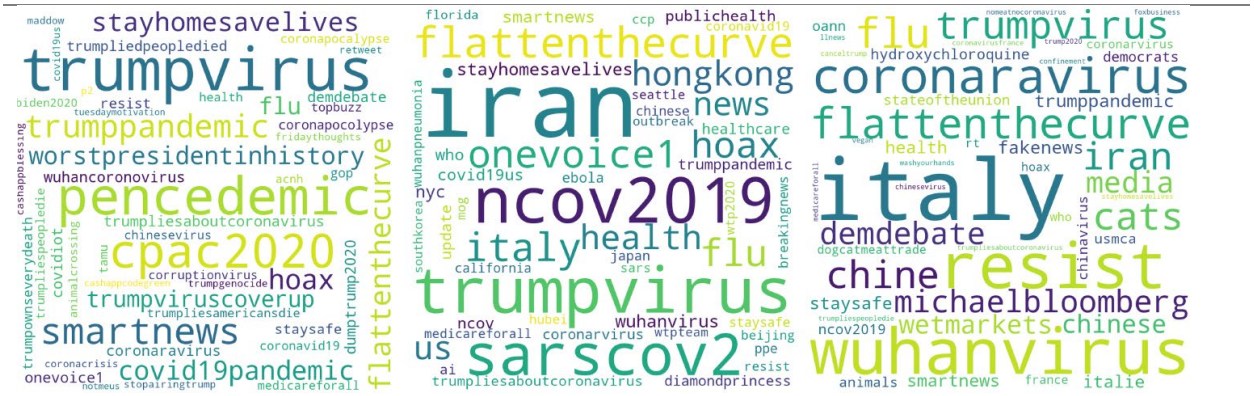

Figure 2: Topics of discussion within the six ideological groups.

We also used this data to study the complexity of polarization. We develop methods to classify the ideological alignment of users along the *moderacy* (hardline vs moderate), *political* (liberal vs conservative) and *science* (anti-science vs pro-science) dimensions. We demonstrate that polarization along the science and political dimensions are correlated, and politically moderate users are more likely to be aligned with the pro-science views, and politically hardline users with anti-science views. Figure 2 shows the common keywords in the messages posted by more than 2 million Twitter users, after they were automatically assigned to different polarized groups by the algorithm we developed. Conspiracy

topics (qanon, wwg1wga) are prevalent on the right, as are topics related to political campaigns. Although the left users talk about “hoax”, they do not mention QAnon conspiracies.

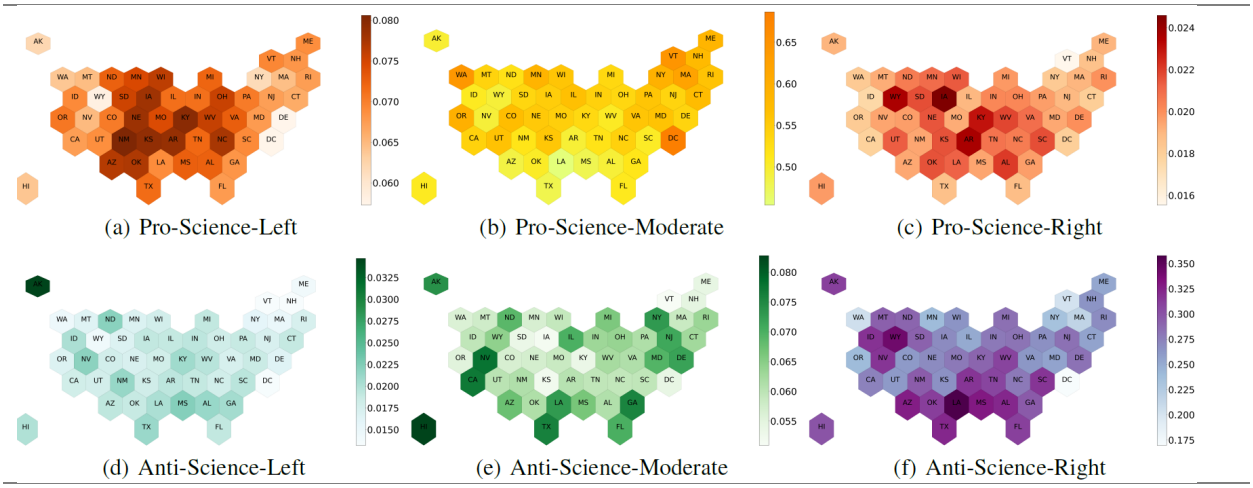

Figure 3: Fraction of state’s Twitter users per ideological category. Figures (a)-(c) show the fraction of states’ Twitter users who are classified as Pro-Science Left, Pro-Science Moderate and Pro-Science Right, espectively. Figures (d)-(f) show the fraction of states’ Twitter users who are classified as Anti-Science Left, Anti-Science Moderate and Anti-Science Right, respectively.

Contrary to expectations, we do not find that polarization grows over time; instead, we see increasing activity by moderate pro-science users. We also show that anti-science conservatives tend to tweet from the Southern US, while anti-science moderates from the Western states (Fig. 2). Our findings shed light on the multi-dimensional nature of polarization, and the feasibility of tracking polarized opinions about the pandemic across time and space through social media data. This paper is under review at the International Conference on the Web and Social Media. We are building on this work to examine how the structure of online networks interacts with opinions. This will allow us to empirically measure how echo chambers of partisan opinion evolve over time.

To better understand the sociological basis of polarization, we applied the same methodology to explore anti-science views expressed by Twitter users in October 2016. By linking tweets to counties in the US via their coordinates, this data allowed us also to study how attitudes towards science relate to socioeconomic characteristics of places from which people tweet. Our analysis revealed three types of places with distinct behaviors: large urban centers, smaller metropolitan regions, and rural areas. Statistical analysis showed that while political partisanship (share of Trump voters) and race (share of White population) are strongly associated with the share of anti-science users across all counties, income was negatively (resp. positively) associated with anti-science attitudes in suburban (resp. rural) areas. Surprisingly, education (share of residents with college degree) did not play an important role in explaining the prevalence of anti-science views in a community. On the other hand, emotions in tweets, specifically negative affect and high arousal (i.e., anger), are expressed in suburban and rural counties with many anti-science users, but not in large urban counties. *This work shows that science skepticism was rampant in 2016, creating ripe conditions for misinformation to spread during the Covid-19 pandemic.*

References

1. Addawood, A., Badawy, A., Lerman, K., & Ferrara, E. (2019, July). Linguistic Cues to Deception: Identifying Political Trolls on Social Media. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 13, No. 01, pp. 15-25).
2. Badawy, A., Lerman, K., & Ferrara, E. (2019, May). Who Falls for Online Political Manipulation?. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 162-168). ACM.
3. Badawy, A., Addawood, A., Lerman, K., & Ferrara, E. (2019). Characterizing the 2016 Russian IRA influence campaign. Social Network Analysis and Mining, 9(1), 31.
4. Badawy, Ferrara, Lerman (2018) "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign" in Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining.
5. L Luceri, A Deb, A Badawy, E Ferrara, Red Bots Do It Better: Comparative Analysis of Social Bot Partisan Behavior. Companion Proceedings of The 2019 World Wide Web Conference, page 1007-1012, 2019.
6. M Stella, E Ferrara, M De Domenico, Bots increase exposure to negative and inflammatory content in online social systems. Proceedings of the National Academy of Sciences Dec 2018, 115 (49) 12435-12440; DOI: 10.1073/pnas.1803470115
7. KC Yang, O Varol, C A. Davis, E Ferrara, A Flammini, F Menczer. Arming the public with artificial intelligence to counter social bots. Human Behavior and Emerging Technologies. 2019 https://doi.org/10.1002/hbe2.115
8. Luceri, Luca et al. Evolution of bot and human behavior during elections. *First Monday*. 2019. doi:https://doi.org/10.5210/fm.v24i9.10213.
9. E. Ferrara, The History of Digital Spam. *Communications of the ACM*, 2019, Vol. 62 No. 8, Pages 82-91. 10.1145/3299768
10. S Kudugunta, E Ferrara (2018) Deep Neural Networks for Bot Detection Information Sciences 467 (October), 312-322.
11. JP Allem, E Ferrara (2018) Could social bots pose a threat to public health? American journal of public health 108 (8), 1005.
12. R Dutt, A Deb, E Ferrara. "Senator, We Sell Ads": Analysis of the 2016 Russian Facebook Ads Campaign. Third International Conference on Intelligent Information Technologies (ICIIT 2018).
13. A Deb, A Majmundar, S Seo, A Matsui, R Tandon, S Yan, JP Allem, E Ferrara. (2018) Social Bots for Online Public Health Interventions. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*.
14. E. Ferrara (2017) Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election First Monday 22(8)
15. Alipourfard, N., Nettasinghe, B., Abeliuk, A., Krishnamurthy, V., & Lerman, K. (2020). Friendship paradox biases perceptions in directed networks. *Nature communications*, *11*(1), 1-9. https://www.nature.com/articles/s41467-020-14394-x
16. Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. JMIR Public Health and Surveillance, 6(2), e19273. https://publichealth.jmir.org/2020/2/e19273
17. E Ferrara (2020) What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday*. https://firstmonday.org/ojs/index.php/fm/article/download/10633/9548
18. Jiang, J., Chen, E., Yan, S., Lerman, K., & Ferrara, E. (2020). Political polarization drives online conversations about COVID-19 in the United States. Human Behavior and Emerging Technologies, 2(3), 200-211. https://onlinelibrary.wiley.com/doi/full/10.1002/hbe2.202

19. Ngo, S. C., Percus, A. G., Burghardt, K., & Lerman, K. (2020). The transsortative structure of networks. *Proceedings of the Royal Society A*, *476*(2237), 20190772. https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2019.0772
20. Pozzana, I., & Ferrara, E. (2020). Measuring bot and human behavioral dynamics. *Frontiers in Physics*, *8*, 125.
21. Ashwin Rao et al. (2021) sMeasuring Multi-Dimensional Polarization in Online Discussions about COVID-19, to appear in Journal of Medical Internet Research.
22. Minda Hu, Ashwin Rao, Kristina Lerman, "Socioeconomic correlates of science skepticism in the US." Under review in *Future Internet*.