Disinformation: Detect to Disrupt*

Craig Corcoran [1] craig@newknowledge.com
Renee DiResta [1] renee@newknowledge.com
David Morar [2] davidcristianmorar@gmail.com
Numa Dhamani [1] numa@newknowledge.com
David Sullivan [1] dave@newknowledge.com
Jeffrey Gleason [1] jeffrey.gleason@newknowledge.com
Paul Azunre [3] azunre@algorine.com
Steve Kramer [1] steve@newknowledge.com
Rebecca Ruppel [1] becky@newknowledge.com

1 Introduction

Disinformation is a long-established psychological manipulation technique that has undergone a technological upgrade in the era of social networks (Jensen et al., 2019). Current major social media platforms have become a vector for various actors to disseminate propaganda and execute disinformation campaigns at scale with the goal of influencing elections (Inkster, 2016), targeting industries and brands (Berthon et al., 2018; Visentin et al., 2019), and acting as agents of polarization, radicalization, and social division (DiResta et al., 2018; Rowe and Saif, 2016). Algorithms optimized for user engagement are now leveraged to influence the growing quantity of people who spend an increasing amount of time on these platforms (Del Vicario et al., 2016). Since correcting false narratives is exceedingly difficult, the ability to detect malign influence operations *before* they achieve mass reach is essential to mitigating their impact. Starting from a general definition of the problem space, we discuss several facets of disinformation campaigns, and then use those properties to formulate quantitative methods for detecting and understanding them.

2 Defining Disinformation

The first key property that defines disinformation is the **intent to influence** the target's opinion or behavior (Wardle and Derakhshan, 2017). Perpetrators of disinformation campaigns leverage deception in an attempt to shift attitudes, or inspire action. To achieve the desired influence, these campaigns use features of the information ecosystem (e.g., ease of creating a false identity) to exploit biases and heuristics in human cognition, including the use of authority, familiarity, and perceived consensus as proxies for truth.

Another distinct characteristic of disinformation is the **intent to deceive** the target regarding the provenance, prevalence, or authenticity of a narrative (Wardle and Derakhshan, 2017). On social media platforms, actors can: 1) create misattributed, false, or manipulated content, 2) use inauthentic accounts to disguise the origin of a narrative and/or the identity of those who wish to spread it, and 3) coordinate or automate factions to create the perception of widespread consensus around a particular topic.

Background and Tactical Summary

Disinformation strategies have evolved since the Cold War to take advantage of the latest and most widely used information technologies, but the goal of manipulating the media and citizens of a targeted population remains largely unchanged (Posetti and Matthews, 2018). Disinformation purveyors—which include state actors, ideologues, mercenaries, trolling factions, and spammers (Ferrara, 2017)—now leverage a far more direct connection to their audience via online community structures, algorithmic dissemination tools, and user-targeting capabilities afforded by social networking platforms. As social platforms democratized content creation, they enabled a proliferation of information sources including a multitude of small media properties; disinformation purveyors have proven themselves adept at hiding within this "new media" environment by masquerading as independent media (Faris et al., 2017). Algorithmic dissemination has afforded a significant increase in the velocity and virality of information transmission (Jensen et al., 2019). In addition, malign actors can exploit anonymity and online identity norms with relative ease, creating fabricated identities that mimic those of a targeted community.

* To be presented at the 2019 Comparative Approaches to Disinformation Workshop at Harvard University.
[1] New Knowledge, Austin, Texas, USA
[2] The George Washington University, Washington, DC, USA
[3] Algorine Inc., Austin, Texas, USA

3 Disinformation Campaign Detection

We outline a computational framework that detects characteristic tactics of disinformation campaigns by tracking the media being propagated (*content*), the networks of accounts involved (*voice*), and the flow of information within and across platforms (*dissemination*). We assert that a comprehensive analysis of all three is required for detecting potential disinformation campaigns.

Prior Work

There are a number of computational approaches that aim to automatically identify disinformation (or misinformation), but most limit their scope to one aspect of the problem (content, voice, or dissemination) (Pérez-Rosas et al., 2018), are designed to operate within the confines of a single platform (e.g., bot detection tailored exclusively toward Twitter) (Davis et al., 2016; Ratkiewicz et al., 2011), and rely on manually labeled training data (Pérez-Rosas et al., 2018; Castillo et al., 2011).

In contrast, our work focuses on providing a human analyst with the context necessary to understand the evolving tactics of disinformation campaigns by jointly analyzing all three aspects in a cross-platform setting. Additionally, our methods dont require labeled training data and are highly scalable, which mitigates the risk of bias introduced by manually labeled data and targeted data collection allowing them to easily be applied to new or dynamic environments.

The Detect-to-Disrupt Framework

Our framework develops narrative- and language-agnostic flags to track the flow of content through networks of accounts and highlight indicators of potential disinformation campaigns. We look for sub-networks that appear to be coordinating, rather than focusing on the credibility of a single account or provenance of a piece of content. We characterize this approach to disinformation detection as a **data funnel**, where each step described in the operates on data filtered through the previous steps.

The Detection Process

Detect Anomalous Content

- Augment content with flags marking the presence of particular content fragments (e.g., images, urls, hashtags, tagged usernames, and snippets of text) enabling observation of the flow of information among—and measure the similarity between—accounts based on what they publish.
- Look for content that is statistically extreme (e.g., anomalously high volume of similar content) to determine *what* content is most relevant.
- Analyze the frequency of content over time for anomalous activity to provide insight into *when* a potential disinformation campaign is most active.

Detect Anomalous Voice

- Construct a cross-platform account network graph that encodes multiple types of relationships, including measures of behavioral similarity (e.g., posting similar content at similar times) and platform interactions (e.g., follow, friend, like, or reply).
- Examine graph anomalies, including highly connected sub-communities and bridges between them, for indicators of who may be part of an organized, intentional faction.

Detect Anomalous Dissemination

- Use content flags to track the propagation of information across the account network, and infer an information flow graph.
- Inspect the information flow graph to understand *how* a disinformation campaign is operating, its tactics (e.g., targeting influencers), and the roles of the accounts involved (e.g., content generators or amplifiers).

Facilitate Analyst Review

- Present the results from each phase of the detection process to the analyst to inform assessments about impact, intent, and attribution.

4 Conclusion

With a thorough knowledge of tactics and strategies in aggregate, platforms and disinformation analysts are better equipped to design relevant interventions to disrupt and mitigate impact.

Early detection is a key component of having the ability to intercept and disrupt disinformation campaigns before they can affect their target audience. The novel cross-platform detection framework we have proposed has the potential to significantly improve our capability to detect disinformation campaigns.

References

P Berthon, E Treen, and L Pitt. 2018. How truthiness, fake news and post-fact endanger brands and what to do about it. *GfK Marketing Intelligence Review*.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA. ACM.

Clayton A. Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. *CoRR*, abs/1602.00975.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proc. Natl. Acad. Sci. U. S. A.*, 113(3):554–559.

Renee DiResta, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. 2018. *The Tactics & Tropes of the Internet Research Agency*. New Knowledge.

Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election.

Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday*, 22(8).

Nigel Inkster. 2016. Information warfare and the us presidential election. *Survival*, 58(5):23–32.

Benjamin Jensen, Brandon Valeriano, and Ryan Maness. 2019. Fancy bears and digital trolls: Cyber strategy with a russian twist. *Journal of Strategic Studies*, 42(2):212–234.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Julie Posetti and Alice Matthews. 2018. A short guide to the history offake news and disinformation. *International Center for Journalists*, pages 2018–07.

Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 249–252, New York, NY, USA. ACM.

Matthew Rowe and Hassan Saif. 2016. Mining pro-isis radicalisation signals from social media users. In *ICWSM-16: 10th International AAAI Conference on Web and Social Media*, pages 329–338.

Marco Visentin, Gabriele Pizzi, and Marco Pichierri. 2019. Fake news, real problems for brands: The impact of content truthfulness and source credibility on consumers' behavioral intentions toward the advertised brands. *Journal of Interactive Marketing*, 45:99 – 112.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe Report*, 27.