# DISINFORMATION: DETECT TO DISRUPT

**Craig Corcoran** [1] **Renee DiResta** [1] **David Morar** [1] **Garrett Honke** [2] **David Sullivan** [1]

**Numa Dhamani** [1] **Jeffrey Gleason** [1] **Paul Azunre** [3] **Steve Kramer** [1] **Becky Ruppel** [1]

## 1 Introduction

Disinformation is a long-established psychological manipulation technique that has undergone a technological upgrade in the era of social networks. 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, targeting industries and brands, and acting as agents of polarization, radicalization, and social division. These campaigns are deployed globally by multiple types of actors, so any attempt to address the problem must be agnostic to region and language. Algorithms optimized for user engagement are now leveraged to influence the growing quantity of people who spend an increasing quantity of time on these platforms. Since correcting false narratives is exceedingly difficult, the ability to detect malign influence operations *before* they achieve mass reach is essential for mitigating their impact. Starting from a general definition of the problem space, we discuss several facets of disinformation campaigns, then we use those properties to formulate quantitative methods for detecting and understanding them. The detection methods' holistic interpretation of disinformation allows for a region-, country- and language-agnostic perspective.

## 2 Defining Disinformation

Disinformation is distinct from other types of misleading information in its **intent to influence** the target's opinion or behavior, and its **intent to deceive** the target regarding the provenance, prevalence, or authenticity of the narrative.

The first key property that defines disinformation is the **intent to influence**. 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.

On social media platforms, actors with the **intent to deceive** can create misattributed, false, or manipulated content, use inauthentic accounts to disguise the origin of a narrative or the identity of those who wish to spread it, or use coordinated factions or automation 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. Disinformation purveyors - which include state actors, ideologues, mercenaries, trolling factions, and spammers - 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. Algorithmic dissemination has afforded a significant increase in the velocity and virality of information transmission. In addition, malign actors can exploit anonymity and online identity norms with relative ease, creating fabricated identities that mimic those of a targeted community.

[1] New Knowledge, Austin, Texas, USA

[2] Watson School of Engineering and Applied Science and the Department of Psychology: Cognitive and Brain Sciences, Binghamton University (SUNY), Binghamton, New York, USA

[3] Algorine Inc., Austin, Texas, USA

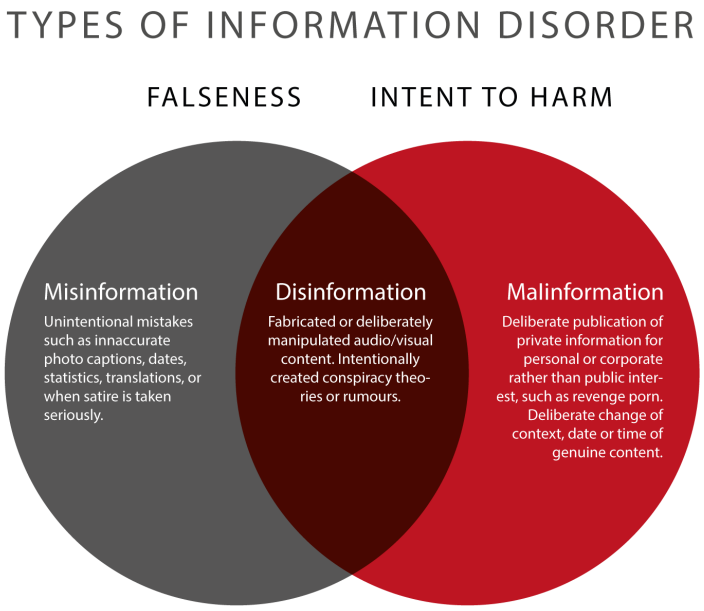


Figure 1: 3 Types of Information Disorder. Credit: Claire Wardle & Hossein Derakshan, 2017.

## 3 Disinformation Campaign Detection

We outline a computational framework that detects characteristic tactics of disinformation campaigns by tracking the media being propagated, the networks of accounts involved, and the flow of information within and across platforms. We refer to these three aspects of a disinformation campaign as content, voice, and 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), are designed to operate within the confines of a single platform (e.g., bot detection tailored exclusively toward Twitter), and rely on manually labeled training data.

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 don't require labeled training data and are highly scalable, which mitigates the risk of bias introduced by manually labeled data and targeted data collection and allows 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 subnetworks that appear to be coordinating, rather than focusing on the credibility of a single account or provenance of a piece of content.

Our approach to disinformation detection can be characterized as a data funnel, where we first use light-weight, scalable algorithms to analyze large amounts of data and identify potentially anomalous content. This step identifies emerging trends, authentic or not - to assess whether a disinformation campaign is involved, we must also examine voice and dissemination. We explore voice by looking at the accounts propagating the emerging content for patterns of behavior that suggest coordinated manipulation. Finally, we look at how information is being disseminated across that network for patterns characteristic of a disinformation campaign. This collection of analyses is presented to a human analyst, who can then qualitatively assess the potential presence of a coordinated, malign campaign meriting additional review or collaborative outside validation.

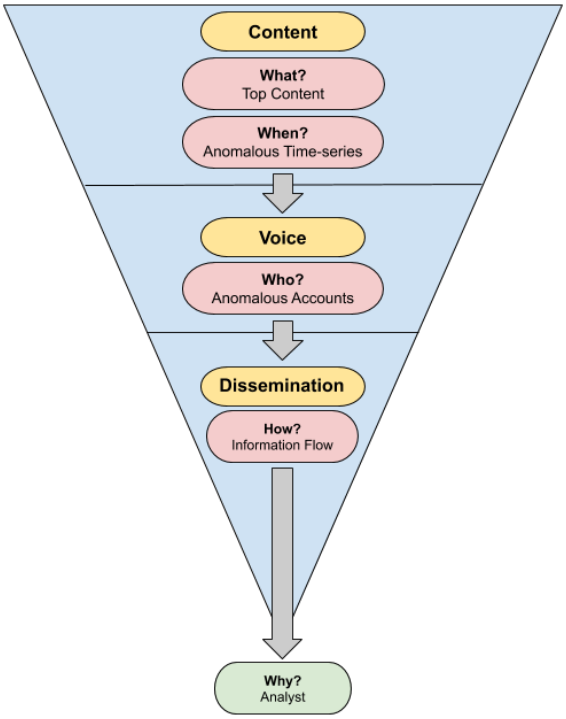


Figure 2: The data funnel detection process.

### 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 subcommunities 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 Moving Forward

With a thorough knowledge of tactics and strategies in aggregate, platforms and counter-campaigners are better equipped to design relevant interventions to disrupt and mitigate impact. The novel detection framework we present has the potential to significantly improve our capability to detect disinformation campaigns across text, image, and video-based social media platforms. With a thorough knowledge of tactics, patterns, and strategies in aggregate, platforms and counter-campaigners are better equipped to design relevant interventions to disrupt and mitigate the impact of disinformation campaigns.

## Author Biographical Notes

**Paul Azunre** (azunre@gmail.com) is an AI Researcher who holds a PhD in Electrical Engineering and Computer Science from MIT, and presently serves as Founder at Algorine Inc.

**Craig Corcoran** (craig@newknowledge.com) is a Data Scientist at New Knowledge who focuses on computational tools for combating disinformation. He holds bachelors degrees in Cognitive Science and Mechanical Engineering from Rice University and studied Machine Learning in the Computer Science PhD program at the University of Texas at Austin (ABD).

**Numa Dhamani** (numa@newknowledge.com) is a Machine Learning Engineer at New Knowledge with a focus on natural language processing and network analysis. She holds Bachelor degrees in Physics and Chemistry from The University of Texas at Austin.

**Renée DiResta** (renee@newknowledge.com) is a 2019 Mozilla Fellow in Media, Misinformation, and Trust. She investigates the spread of malign narratives across social networks, and assists policymakers in understanding and responding to the problem. She has advised Congress, the State Department, and other academic, civic, and business organizations, and has studied disinformation and computational propaganda in the context of pseudoscience conspiracies, terrorism, and state-sponsored information warfare.

**Jeffrey Gleason** (jeffrey.gleason@newknowledge.com) is a Jr. Machine Learning Engineer at New Knowledge, where he focuses on natural language processing, network analysis, time series anomaly detection and algorithmic fairness. He got his degree in Computer Science from Princeton University and wrote a thesis about risk assessment algorithms in the criminal justice system.

**Garrett Honke** (ghonke1@binghamton.edu) is a Machine Learning Research Scientist with a PhD in Cognitive Neuroscience from Binghamton University.

**Steve Kramer** (steve@newknowledge.com) is a Senior Research Scientist at New Knowledge with a background in network graph analysis, natural language processing, anomaly detection, machine learning, and complex systems. Steve applied techniques from chaos theory and nonlinear dynamics to create a patented dynamic anomaly technology to find the "unknown unknowns" in multiple types of time-dependent data sets. Since 2011, he has served as a reviewer and program committee member for the ACM KDD and IEEE Security and Intelligence Informatics conferences. He holds a PhD in Computational Physics from University of Texas at Austin.

**David Morar** (david.morar@newknowledge.com) is a technology policy scholar, most recently a Policy Manager with New Knowledge. Dr. Morar's work has appeared in peer-reviewed conference proceedings, journals, as well as in popular press. He holds a PhD in Public Policy from George Mason University, a Masters of International Affairs from Penn State University and a Bachelors of Political Science from the University of Bucharest.

**Becky Ruppel** (becky@newknowledge.com) is a Data Scientist with New Knowledge and has a Masters degree in Biology from Syracuse University and a Bachelors of Science in Ecology and Evolutionary Biology from the University of Arizona. During her Bachelors and Masters she studied ecology and population genetics of native bees across large geographic ranges to improve species conservation practices. Her work on native bees has been published in peer reviewed journals and she was a co-author on the report for the Senate Select Committee on Intelligence, The Tactics and Tropes of the Internet Research Agency.

**David Sullivan** (dave@newknowledge.com) is a Data Scientist with New Knowledge and holds a Masters Degree in Computer Science specializing in Machine Learning and a Bachelor's Degree in Physics, both from Rice University. He is also a co-author on the report for the Senate Select Committee on Intelligence, The Tactics and Tropes of the Internet Research Agency.