EIC'S MESSAGE

EDITOR IN CHIEF JEFFREY VOAS
IEEE Fellow; j.voas@ieee.org

"Propagating" Disinformation

Jeffrey Voas, IEEE Fellow

This message ponders whether traditional propagation analysis techniques from the reliability and safety communities can offer any insight into how to thwart the propagation of disinformation.

A lie is a lie. A half-truth is a half-truth. Is an omission of information a half-truth or a lie? And what about misinformation and disinformation? There is a spectrum between lies and absolute truth. Answering these questions (and others) affects our ability to trust.

I asked ChatGPT two simple questions:

1. How does Internet information go viral? ChatGPT gave eight criteria: shareability, emotional appeal, timeliness and relevance, uniqueness and novelty, simplicity and accessibility, influencers and networks, seeding and initial exposure, and user engagement and participation.
2. How does disinformation spread on Twitter? ChatGPT gave nine criteria: bots, coordinated campaigns (multiple accounts working together), hashtag hijacking (making hashtags to promote their disinformation), fake accounts, amplification by influential users, confirmation bias (engage with content that aligns with their existing beliefs and biases), lack of fact-checking, clickbait, and retweeting without reading.

I've spent years analyzing how corrupted internal data states propagate while "a piece of software" executes. Part of those efforts resulted in automated prototype tools for both source code and off-the-shelf software packages (black boxes with access only to the interfaces).[1] These tools injected artificial, corrupt data into executing software and then observed what effect the corrupted data had on the software's output behavior.

Propagation of corrupted data during software and system operation is a problem that disciplines such as fault tolerance, reliability, and dependability (FRD) seek to mitigate, particularly if the propagation results in hazardous outcomes. To perform such mitigation, the software

DISCLAIMER

The author is completely responsible for the content in this message. The opinions expressed here are his.

Digital Object Identifier 10.1109/MC.2023.3297809
Date of current version: 13 November 2023

and system must be bounded. During normal operation, corrupted data can result from both malicious intent and nonmalicious faults, as well other data-corrupting enablers (e.g., simply reading in corrupt data from a sensor).

Information comes in many forms and often is textual data; disinformation is data that results from malicious intent. Social media is often labeled as a source of disinformation. So, are there current automated FRD techniques to thwart data propagation that could be applied to social media platforms? However, like the Internet, social media platforms are far from bounded. The Internet and blogosphere are firehoses of information.

There is a spectrum between lies and absolute truth.

It is curious to consider whether any of the FRD disciplines can be used to halt or reduce the propagation of disinformation. If we injected artificial disinformation into a social media platform, could we see how it propagates and is republished? Might we create a new problem like the disinformation about the Titan submersible (https://www.newsweek.com/titan-submersible-implosion-screams-tiktok-1811033)? Or could we experiment in a controlled/bounded (laboratory-like) setting, and if so, how?

Recognize the difference between a fixed (bounded) *software system* with instrumented break points designed to halt a "bad" execution (or return it to a safe state) and a *social media platform*. Bounded systems have interconnections that can be automatically traced. But unbounded social media platforms have human participants that are not statically connected; people (and their information) come and go instantly (just like the Internet that is continuously changing). However, "friending" and "following" may allow for static tracing of human interconnections.

In Twitter, information can quickly be retweeted, making it viral. Current processes for stopping disinformation propagation in Twitter and Facebook are mostly manual from what is reported; the mainstream media (the fourth estate) is often the first to sound an alarm. One approach used for thwarting the dissemination of disinformation by social media vendors is to ban or lock out certain people, one by one. This is often applied to high-profile individuals (influencers). So, is a "locking out" approach tractable for millions of people? Of course not. Another approach is for social media providers to reactively and manually pull out disinformation, but this is probably intractable and only occurs once the social media provider notices the situation and verifies that the information is indeed malicious. And what about those *fact checkers* that need real-time execution speed and still create false positives and negatives due to inaccuracy? Can we take corrective actions to disable disinformation quickly enough given current fallibility of fact checking approaches? No.

In closing, social media providers may already have proprietary, automated tools to study information propagation. Applying automated tools to inject artificial disinformation in a controlled environment to study propagation may be a reasonable avenue for research considering that the U.S. Government is already focused on thwarting the dissemination of disinformation from adversarial nation states (https://www.state.gov/disarming-disinformation/).

REFERENCE

1. J. Voas, F. Charron, G. McGraw, K. Miller, and M. Friedman, "Predicting how badly 'Good' software can behave," *IEEE Softw.*, vol. 14, no. 4, pp. 73–83, Jul./Aug. 1997, doi: 10.1109/52.595959.

JEFFREY VOAS, Gaithersburg, MD 20899 USA, is the editor in chief of *Computer*. He is a Fellow of IEEE. Contact him at j.voas@ieee.org.