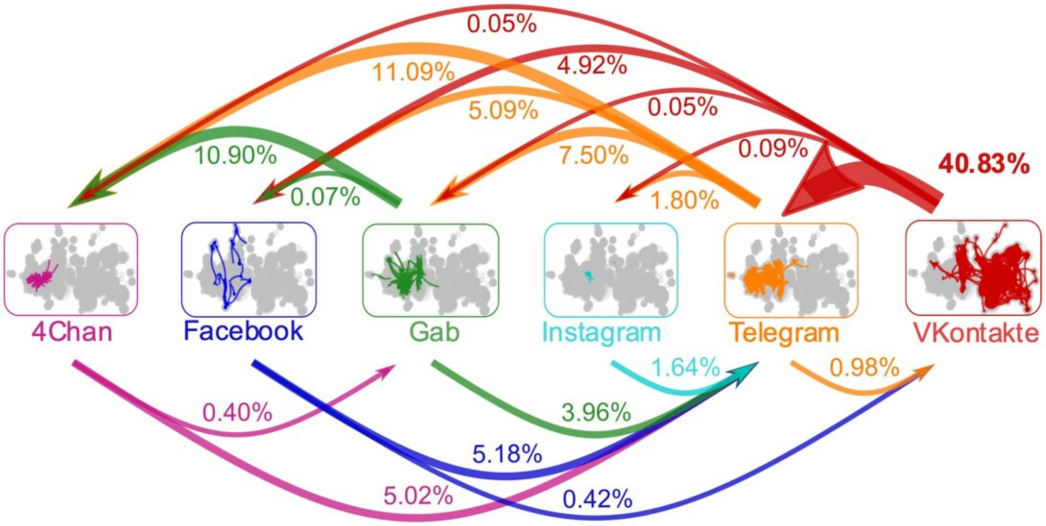
PHYS.ORG

Malicious content exploits pathways between platforms to thrive online, subvert moderation

June 15 2021



Malicious COVID-19 content (e.g. anti-Asian hate) exploits pathways between social media platforms to spread online. Credit: Neil Johnson/GW

Malicious COVID-19 online content—including racist content, disinformation and misinformation—thrives and spreads online by bypassing the moderation efforts of individual social media platforms, according to new research published in the journal *Scientific Reports*. By mapping online hate clusters across six major social media platforms, researchers at the George Washington University show how malicious content exploits pathways between platforms, highlighting the need for social media companies to rethink and adjust their content moderation policies.

Led by Neil Johnson, a professor of physics at GW, the research team set out to understand how and why malicious content thrives so well online despite significant moderation efforts, and how it can be stopped. The team used a combination of machine learning and network data science to investigate how online hate communities sharpened COVID-19 as a weapon and used current events to draw in new followers.

"Until now, slowing the spread of malicious content online has been like playing a game of whack-a-mole, because a map of the online hate multiverse did not exist," Johnson, who is also a researcher at the GW Institute for Data, Democracy & Politics, said. "You cannot win a battle if you don't have a map of the battlefield. In our study, we laid out a first-of-its-kind map of this battlefield. Whether you're looking at traditional hate topics, such as anti-Semitism or anti-Asian racism surrounding COVID-19, the battlefield map is the same. And it is this map of links within and between platforms that is the missing piece in understanding how we can slow or stop the spread of online hate content."

The researchers began by mapping how hate clusters interconnect to spread their narratives across social media platforms. Focusing on six platforms—Facebook, VKontakte, Instagram, Gab, Telegram and 4Chan—the team started with a given hate cluster and looked outward to find a second cluster that was strongly connected to the original. They found the strongest connections were VKontakte into Telegram (40.83% of cross-platform connections), Telegram into 4Chan (11.09%), and Gab into 4Chan (10.90%).

The researchers then turned their attention to identifying malicious content related to COVID-19. They found that the coherence of COVID-19 discussion increased rapidly in the early phases of the pandemic, with hate clusters forming narratives and cohering around COVID-19 topics and misinformation. To subvert moderation efforts by social media platforms, groups sending hate messages used several adaptation strategies in order to regroup on other platforms and/or reenter a platform, the researchers found. For example, clusters frequently change their names to avoid detection by moderators' algorithms, such as vaccine to va$$ine. Similarly, anti-Semitic and anti-LGBTQ clusters simply add strings of 1's or A's before their name.

"Because the number of independent social media platforms is growing, these hate-generating clusters are very likely to strengthen and expand their interconnections via new links, and will likely exploit new platforms which lie beyond the reach of the U.S. and other Western nations' jurisdictions." Johnson said. "The chances of getting all social media platforms globally to work together to solve this are very slim. However, our mathematical analysis identifies strategies that platforms can use as a group to effectively slow or block online hate content."

Based on their findings, the team suggests several ways for social media platforms to slow the spread of malicious content:

- Artificially lengthen the pathways that malicious content needs to take between clusters, increasing the chances of its detection by moderators and delaying the spread of time-sensitive material such as weaponized COVID-19 misinformation and violent content.
- Control the size of an online hate cluster's support base by placing a cap on the size of clusters.

- Introduce non-malicious, mainstream content in order to effectively dilute a cluster's focus.

"Our study demonstrates a similarity between the spread of online hate and the spread of a virus," Yonatan Lupu, an associate professor of political science at GW and co-author on the paper, said. "Individual social media platforms have had difficulty controlling the spread of online hate, which mirrors the difficulty individual countries around the world have had in stopping the spread of the COVID-19 virus."

Going forward, Johnson and his team are already using their map and its mathematical modeling to analyze other forms of malicious content—including the weaponization of COVID-19 vaccines in which certain countries are attempting to manipulate mainstream sentiment for nationalistic gains. They are also examining the extent to which single actors, including foreign governments, may play a more influential or controlling role in this space than others.

More information: *Scientific Reports* (2021). DOI: 10.1038/s41598-021-89467 , www.nature.com/articles/s41598-021-89467-y

Provided by George Washington University

Citation: Malicious content exploits pathways between platforms to thrive online, subvert moderation (2021, June 15) retrieved 25 April 2024 from https://phys.org/news/2021-06-malicious-content-exploits-pathways-platforms.html