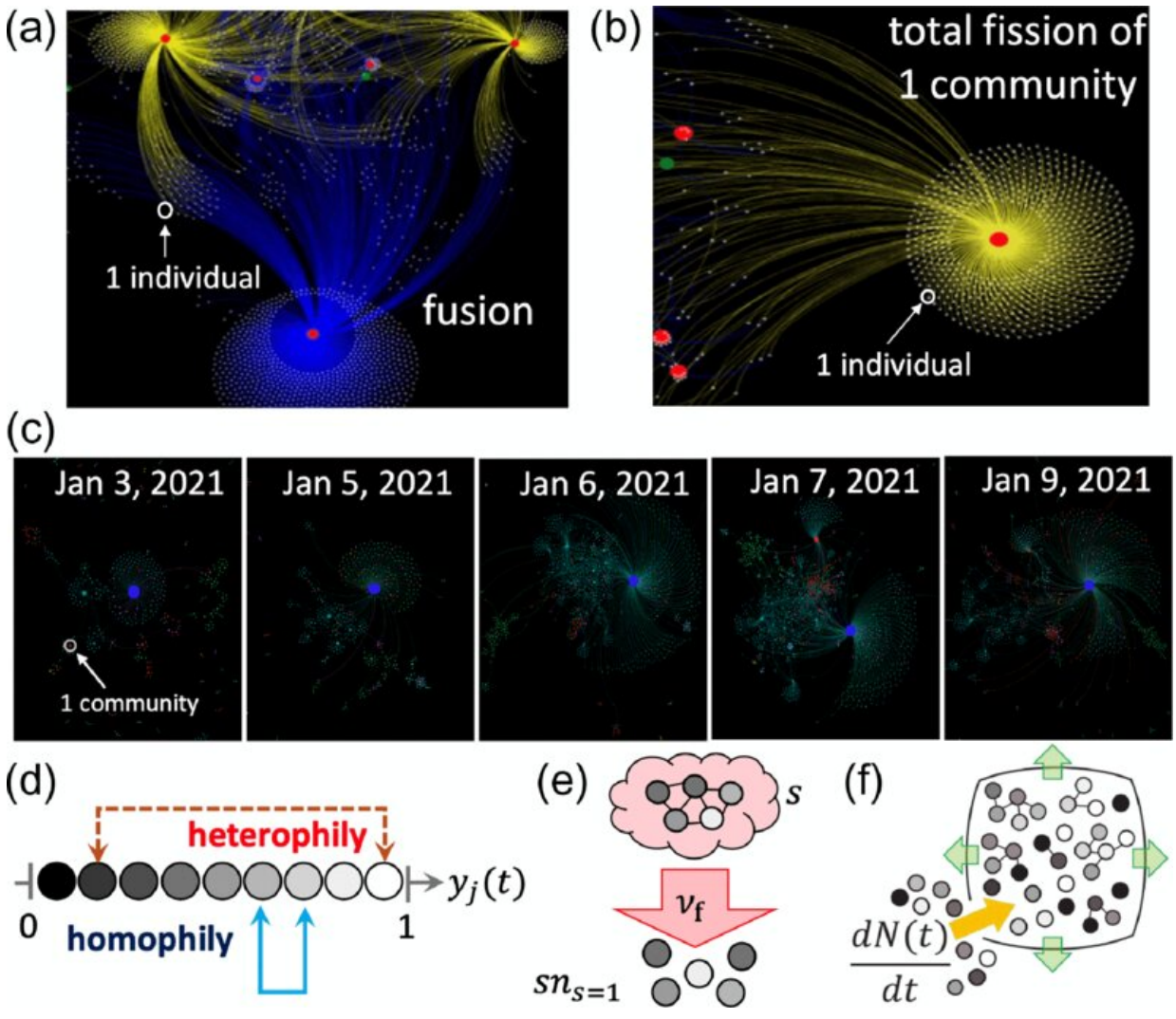
Researchers develop online hate speech 'shockwave' formula

June 6 2023



Empirically observed (a) fusion and (b) total fission of in-built communities featuring anti-U.S. hate on VKontakte between day t (yellow) and t + 1 (blue). Red nodes are anti-U.S. communities that later got shut down (total fission);

green nodes are those still not yet shut down; yellow links point to individuals (white dots) removed from the anti-U.S. community on day t + 1 ; blue links point to individuals added to the anti-U.S. community on day t + 1. Spatial layout results from (a) and (b) being close-ups of a fuller network plotted using ForceAtlas2, meaning that nodes appearing closer together are more interconnected. (b) Also shows that very few individuals are simultaneously also members of other communities (SM shows further proof). (c) Empirically observed clustering of antigovernment communities across platforms around U.S. Capitol riot. (d)–(f) The theory in this Letter incorporates (d) heterogeneous individuals aggregating (i.e., fusion) based on character similarity, (e) total fission with probability ν f , (f) time-varying population size N (t). Credit: *Physical Review Letters* (2023). DOI: 10.1103/PhysRevLett.130.237401

A George Washington University research team has created a novel formula that demonstrates how, why, and when hate speech spreads throughout social media. The researchers put forth a first-principles dynamical theory that explores a new realm of physics in order to represent the shockwave effect created by bigoted content across online communities.

This effect is evident in lightly moderated websites, such as 4Chan, and highly regulated social platforms like Facebook. Furthermore, hate speech ripples through online communities in a pattern that non-hateful content typically does not follow.

The article was published in *Physical Review Letters* on June 5, 2023.

The new theory considers recently gained knowledge on the pivotal role of in-built communities in the growth of online extremism. The formula weighs the competing forces of fusion and fission, accounting for the spontaneous emergence of in-built communities through the absorption of other communities and interested individuals (fusion) and the

disciplinary measures moderators take against users and groups that violate a given platform's rules (fission).

Researchers hope the formula can serve as a tool for moderators to project the shockwave-like spread of hateful content and develop methods to delay, divert, and prevent it from spiraling out of control. The novel theory could also be applied beyond social media platforms and online message boards, potentially powering moderation strategies on blockchain platforms, generative AI, and the metaverse.

"This study presents the missing science of how harms thrive online and, hence, how they can be overcome," Neil Johnson, professor of physics at the George Washington University and co-author of the study, said. "This missing science is a new form of shockwave physics."

More information: Pedro D. Manrique et al, Shockwavelike Behavior across Social Media, *Physical Review Letters* (2023). DOI: 10.1103/PhysRevLett.130.237401

Provided by George Washington University