INVESTIGATION · DIGITAL INVESTIGATION

Investigating Disinformation Campaigns

Disinformation campaigns spread false or misleading content on purpose, often through networks of fake or coordinated accounts. This article explains how researchers and investigators identify such campaigns, trace who is behind them and judge how much effect they had.

ModQu research article · 1,140 words · 13 references (APA 7th edition)

False information has always circulated, but social media allow anyone to publish to large audiences at almost no cost, and allow organised groups to make a message appear far more popular than it is. Investigating disinformation campaigns has therefore become an important task for journalists, academic researchers, platform trust and safety teams and government agencies. The work combines content analysis, network analysis and traditional investigation, and it requires careful separation of what can be proven from what is merely suspected.

Definitions

Wardle and Derakhshan (2017) distinguish three kinds of "information disorder". **Misinformation** is false information shared without intent to harm, such as a mistaken rumour passed on in good faith. **Disinformation** is false information created or shared deliberately to cause harm or gain advantage. **Malinformation** is genuine information shared to cause harm, such as leaked private documents released to damage a person. Investigations of campaigns are mostly concerned with disinformation, and with the organised effort behind it rather than with individual false posts.

A useful concept here is what platforms call coordinated inauthentic behaviour: groups of accounts or pages that work together to mislead people about who they are and what they are doing (Gleicher, 2018). Under this approach, the problem is not only whether content is false, but whether the people spreading it are hiding their identity and coordination. That distinction matters because much effective propaganda uses true but selective facts, and because judging deception about identity is often more objective than judging the truth of contested claims.

Why disinformation spreads

Research helps explain why investigators find so much material. Vosoughi et al. (2018) analysed around 126,000 stories shared on Twitter between 2006 and 2017 and found that false news spread farther, faster and more broadly than true news, largely because people found it more novel. Lazer et al. (2018) reviewed the science of fake news and called for research on both its spread and its effects. Pennycook and Rand (2021) found that people often share false content not because they believe it, but because they do not stop to think about accuracy before sharing. Campaigns exploit these tendencies by producing emotional, surprising and identity-affirming content.

How campaigns are identified

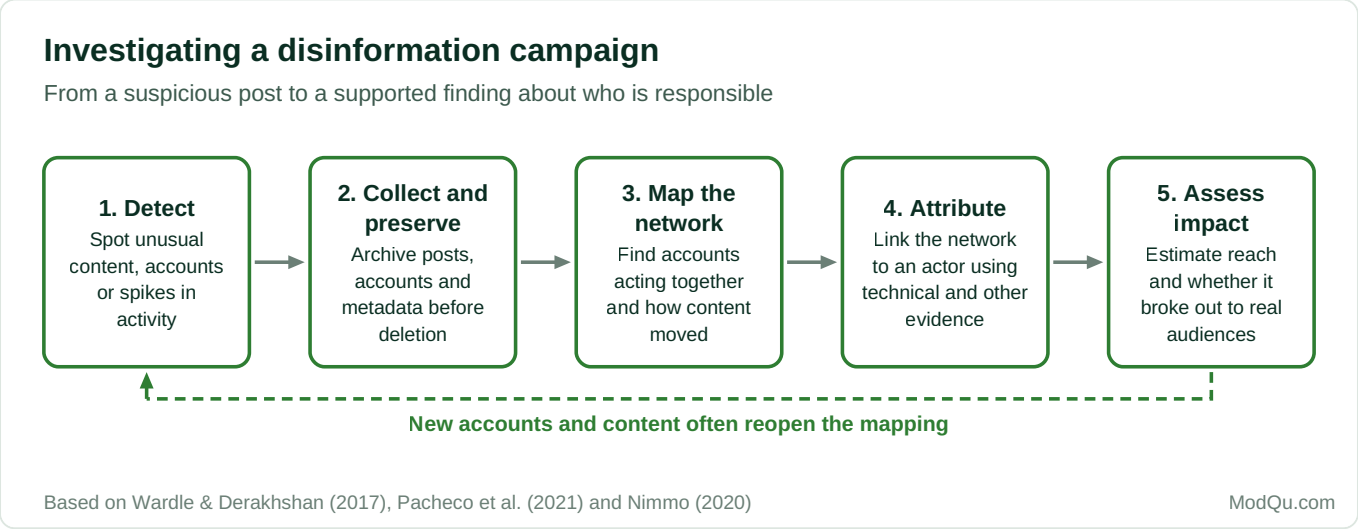


Figure 1. Five stages in investigating a disinformation campaign. Each stage produces evidence with different levels of certainty; attribution is usually the hardest. Diagram by ModQu.

Investigations usually begin with detection: a researcher notices a sudden surge of identical posts, a cluster of new accounts with stock profile pictures, or a story that appears simultaneously on unrelated websites. The next step is preservation. Accounts and posts are often deleted once a campaign is exposed, so investigators archive them, with time stamps and identifiers, before analysis.

The central analytical task is to find coordination. Investigators look for signals such as:

- **Shared content**: many accounts posting the same text, images or links within a short time.
- **Synchronised timing**: accounts that post in bursts at the same moments, or keep working hours in a particular time zone.
- **Account features**: recent creation dates, few original posts, recycled or generated profile photos and similar naming patterns.
- **Amplification structure**: small groups of accounts that consistently share and reply to each other to push content into trending lists.

Pacheco et al. (2021) describe general methods for uncovering such coordinated networks by building networks of accounts linked by shared behaviour and then examining unusually dense clusters. Keller et al. (2020) studied a South Korean intelligence agency's campaign on Twitter and showed that coordination patterns, such as accounts posting the same messages at the same time, revealed the campaign even when individual accounts looked like ordinary users. Ferrara et al. (2016) review how automated accounts, or social bots, can be detected, while noting that many campaigns rely on human operators who are harder to separate from genuine users.

Disinformation is also rarely the work of a few hidden accounts alone. Starbird et al. (2019) show that strategic information operations are collaborative, combining orchestrated accounts with real users who pick up and spread the messages. This makes the boundary of a campaign difficult to draw, and investigators must avoid treating ordinary people who share content as part of the operation.

Attribution

Finding coordinated accounts is not the same as knowing who controls them. Attribution draws on technical evidence that platforms may hold, such as shared devices, internet addresses and payment details, together with content evidence, language patterns, and links to known operations. Public researchers usually have less technical data than platforms and so state attribution with appropriate caution, distinguishing between "linked to", "consistent with" and "operated by".

The best-documented case is the activity of the Internet Research Agency, a company based in St Petersburg, during the 2016 United States presidential election. The investigation led by Robert Mueller set out how the company created fake American personas and groups on social media, bought advertisements and organised rallies, and it led to indictments of individuals and companies (Mueller, 2019). Rid (2020) places such operations in the long history of state "active measures", which have used forged documents and planted stories since well before the internet.

Measuring impact

A campaign that reaches only its own fake accounts has little effect. Nimmo (2020) proposed the Breakout Scale, which rates operations from category one, where content stays within one community on one platform, to category six, where it prompts a policy response or calls for violence. The scale helps investigators avoid both overstating and understating a campaign's importance. Benkler et al. (2018) argue, from a large study of the American media ecosystem, that domestic partisan media often did more to spread false narratives than foreign operations did, a reminder that the visible campaign may not be the main source of a problem.

A short example

The following hypothetical case illustrates the approach. Before a local election, a researcher notices that dozens of accounts share the same claim that one candidate's company polluted a river. Archiving the posts shows that the accounts were created within the same two weeks, use photographs found elsewhere online, and post in identical bursts. Network analysis links them to a handful of websites registered on the same day. The researcher reports that the accounts show strong signs of coordinated inauthentic behaviour, notes that the pollution claim is not supported by environmental records, and states that the people behind the network have not been identified. The platform later removes the accounts, citing the same coordination evidence.

Conclusion

Investigating disinformation requires separating three questions: is the content false or misleading, are the accounts spreading it coordinated and deceptive about their identity, and who is responsible. The first can often be checked against records, the second can be shown through patterns of behaviour, and the third is usually the hardest to prove. Clear statements of what is known, and how confidently, protect both the public and the people wrongly suspected of taking part.

References

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

Gleicher, N. (2018, December 6). *Coordinated inauthentic behavior explained*. Meta Newsroom.

Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2020). Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication*, *37*(2), 256–280. https://doi.org/10.1080/10584609.2019.1661888

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Mueller, R. S., III. (2019). *Report on the investigation into Russian interference in the 2016 presidential election* (Vol. 1). U.S. Department of Justice.

Nimmo, B. (2020). *The Breakout Scale: Measuring the impact of influence operations*. Brookings Institution.

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021). Uncovering coordinated networks on social media: Methods and case studies. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 455–466. https://doi.org/10.1609/icwsm.v15i1.18075

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Rid, T. (2020). *Active measures: The secret history of disinformation and political warfare*. Farrar, Straus and Giroux.

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), Article 127. https://doi.org/10.1145/3359229

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (Council of Europe report DGI(2017)09). Council of Europe.