Combating Online Misinformation Through Game-Theoretic Incentives

An STS Research Project

In STS 4600

Presented to

The Faculty of the

School of Engineering and Applied Science University of Virginia

In Partial Fulfillment of the Requirements for the Degree Bachelor of Science in Computer Science

By Gustavo Moreira

March 21, 2022

Technical Team Members: Jibang Wu

Advisor: Travis Elliott

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: Gustavo Moreira Date: May 15, 2022

Abstract

In recent years, online misinformation has become a major talking point as a result of various high profile cases in which suh misinformation has drastically changed the outcome of various world events, such as the 2020 US presidential election and various policies concerning the prevention of the spread of COVID-19. In response, major social media outlets have tried implementing a variety of techniques to limit the spread of misinformation, such as banning accounts and adding information tags to posts involving controversial topics. However, these attempts have been largely ineffective because they have focused on punishing unwanted behavior rather than incentivizing "good" behavior in the form of posting verifiable and well-sourced information. Using preliminary research results regarding optimal behavior in a Bayesian persuasion scenario, this research project seeks to motivate a policy recommendation for Twitter based on positive reinforcement of truthful posting as a way to combat misinformation

Introduction

In a world in which information is being transmitted at an ever-increasing rate across the internet, particularly over social media apps, it can be difficult to verify what information is based in factual reporting and what information is either intentionally or unintentionally misleading. The latter type of information has come to be known as "misinformation" in discussions around the internet about the phenomenon, and it has been particularly relevant to a variety of different major political events over the past few years, including information about the COVID-19 pandemic and the 2020 presidential election (Chen, Chang, Lerman, Cowan & Ferrara, 2021, p.4). Misinformation has been an especially hot topic of debate with regards to the variety of attempts various social media platforms like Twitter and Facebook have made to limit its spread (Alba, 2021, p.1). Most of these attempts involved limiting the ability of certain users from posting based on their posting history or automatically flagging any posts with certain keywords with sourced information blurbs about the issue at hand (Valenzuela, Halpern, Katz, & Miranda, 2019, p.6). These methods, while effective in some cases, were largely dismissed as ineffective and decried as limiting the freedom of users to post certain information.

Framework Introduction

When analyzing any complex socio-technical problem, an important level of analysis is that of the network of relationships between the different stakeholders. A traditional way to do this in the field of STS is through the use of Actor Network Theory (ANT), which is specifically designed to interpret and understand the relationships between concepts, entities, and individuals that are related to a problem (Crawford, 2020, p.1). Specifically, ANT dictates that all things that exist in the natural world are defined by relationships in an ever evolving and changing network,

which implies that understanding these relationships is sufficient to fully describe how any given problem is constructed. The theory was originally developed in the late 1980s at the Centre de Sociologie de l'Innovation by scholars Michel Callon, Madeleine Akrich, and and Bruno Latour in response to many other sociological theories that did not privilege concepts as being able to act on relationships independent of the actions of human agents, as the belief that semiotic and material relationships can be deeply interconnected was fairly radical at the time.

Typically, the way one constructs an ANT analysis of a given problem is to begin by listing all of the relevant entities to the problem, where an entity in this instance can refer to people, organizations, concepts, and more; essentially any relevant extant thing can be considered part of the network if it is relevant to understanding the problem. Once all of the relevant entities have been addressed, the next key step is to specify the relationships between these entities and how these entities are interconnected. These relationships are not specified by the connection itself, but rather by the concepts to which they relate. For example, if one were to create an ANT graph for a website like YouTube, one would have to make a connection between users and the concepts of "designer" and "creator", as those are just a few of the roles that a user on that platform can take.

Clearly, the process as described above could quickly grow out of hand given the vast possibility of including minute details and conceptual differences for even a simple problem. To that end, it is often beneficial to narrow the scope of the analysis to focus only on the entities that are strictly necessary in order to gain something meaningful from the analysis. For example, in the previously mentioned case of an ANT analysis of YouTube, one could include things like specific content creators in their analysis, as all such creators are affected by YouTube. However, if one is trying to develop policy recommendations for YouTube as a platform, it is unlikely that

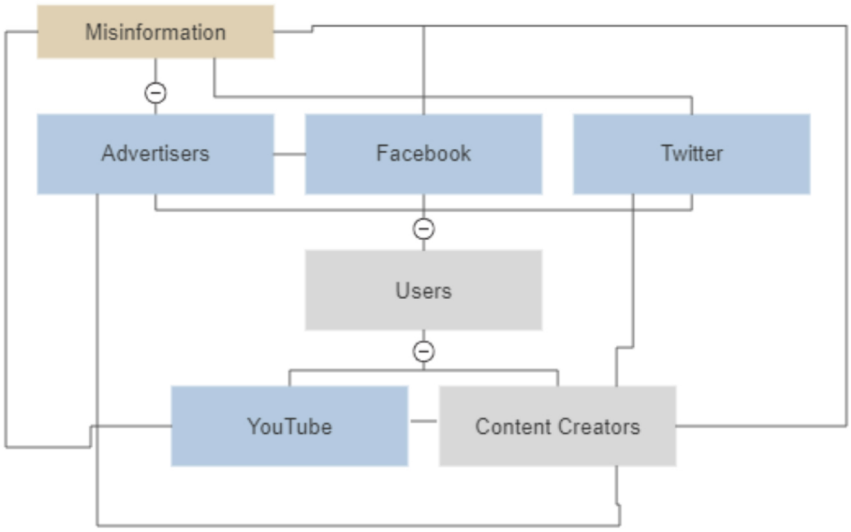
a specific creator is important enough to be included in the analysis, so it makes more sense to leave their connection with the rest of the network as being solely implicit based on those entities that the analyzer does choose to include.

An ANT Analysis of Social Media Misinformation

With regards to the problem of misinformation online (and on social media in particular), an ANT analysis is especially appropriate because there is already a built in network framework between the relevant entities on the platform that can be naturally extended to include concepts and other relevant organizations to complete the analysis. In particular, social media platforms generally are made up of three main groups with some level of intermingling between them: platform hosts, consumers, and content creators. On platforms like Twitter, users and content creators have almost identical abilities, while on YouTube there are more clear delineating ability differences between the two groups. This sort of mixing between the groups can make parsing the connections between them more challenging, but by and large most social media platforms have a clear distinction in the different roles users can take as either content creators or consumers. Key to both user roles is the idea of sharing, as ultimately the principle goal of most social media platforms is to share information about users' lives with other users, and this same notion of sharing is where much of the core problems of misinformation begin.

We now transition our primary focus of discussion to misinformation spread across Twitter in particular as a platform, as Twitter is for many people used as a major news source and has been the site of several key misinformation and censorship controversies over the past few years, especially regarding Donald Trump's removal from the website. By Twitter's very nature of only allowing 240 word posts, it can often be a cesspool of over-simplifications and appeals to

the most radical worldviews possible purely as a way to gain virality. With regards to an ANT analysis, it is clear that the more popular accounts serve as the primary content creators and controllers of what information becomes viral while most regular users simply read posts from those content creators, rarely reaching the level of viral. The platform host, then, is the one that controls what sorts of posts go on the trending page and determines what sorts of created content are sent to users. This is of course computed almost entirely by an algorithmic rather than manual process, but it is nonetheless very much controlled by Twitter as a corporation. This combination of factors leads to the following figure, an ANT network diagram that represents the the various important entities in the social media landscape:



The reason I have included these other social media platforms besides Twitter is that their content does have a relevant effect on what users post on Twitter, as posts are often cross posted to Facebook and embedded YouTube links are very standard. An important other component not yet previously addressed that is shown by this diagram is the role of advertising, as the incentive structure of almost all social media platforms is almost entirely based on gaining as many

advertising dollars as possible, so any policy recommendation made to these platforms needs to be founded in a justification that said policy change will increase long term or short term advertising revenue, which in turn requires that said policy recommendation improves advertising opportunities.

Policy Recommendation: Positive Reinforcement for Truthful Posting

As previously mentioned, almost all previous attempts of mitigating misinformation spread online have been based on punishing those who post misinformation rather than benefiting those that post truthful information. The core of our proposal is to flip this dynamic: create a system in which those who post well-sourced, cited, and nuanced information receive benefits rather than punishing those who post misinformation. This proposal comes as a result of research by the SIGMA lab at UVA on Bayesian Persuasion, a game-theoretic model in which one intelligent agent tries to convince another to act in a certain way to maximize their reward. Our preliminary results suggest that agents that act in cooperative environments (environments where both agents receive a shared reward) are much more likely to converge to an agreed upon set of actions than those operating in a competitive environment (environments where agents receive distinct [and potentially opposite] rewards). With some extrapolation, this suggests that a strategy in which platform revenue is divested among content creators who post in a way desired by the platform may be more successful than a strategy based on punishing malicious users.

A system of rewarding positive behavior can come in a variety of different flavors. The simplest and most straightforward option is to simply pay individuals who post verified political information, where the amount paid is proportional to the reach that the post has and the total amount paid out is some proportion of the company's revenue. However, this would most likely

result in spamming of well known and trivially uncontroversial facts with verified information and would moreover fall into similar issues of necessitating algorithmic verification of sources which can be fairly unreliable. Another initial common sense attempt would be to give every user credits they can award to posts they believe are well-sourced, and give some profit share to those posts that meet a certain award threshold, where the profit is again proportional to the most awarded posts from a set pool taken as part of the company's profits. However, this system too would likely result in abuse, wherein members of rival political factions would coordinate to knowingly give their credits to misinformative posts in order to either anger their opponents or convince those "on the fence", particularly in a political climate as polarized as that of the United States. However, there is an alternative that combines the best of both of these approaches that is more resilient (though not immune) to abuse

The key change this proposal makes to these two naive approaches is to use the positive reinforcement as a filtering system for manual checking with tiers of users rather than requiring algorithmic truth verification. In other words, the system would involve creating three (or more) tiers of users that each have different abilities to monitor misinformation on these social media platforms. This borrows from some of the systems used for open source projects like Wikipedia, where those editors who are known to submit quality work are given additional privileges in a tiered way, where a user can only enter a higher tier via verification by another user in that tier.

Conclusion

Misinformation is a major source of social tension online and offline. The previous attempts that have been made to combat misinformation have been largely ineffective because they misunderstand the incentive structures necessary to ensure that bad actors on these social

media platforms are incentivized to post verifiable information. Using the insights gained from experiments in algorithmic game theory, it is likely that the above proposal to incentivize truthful information posting on Twitter would have a net positive effect on reducing misinformation on the platform and reduce the amount of polarization such platforms cause. This proposal was informed both by these game theoretic results and an analysis of the social media landscape through the use of ANT, which enabled us to determine which agents in this system needed to modify policy in order to incentivize the desired outcome (that being the platform owner, Twitter, in this case). We hope this combination of analytic factors allows others in the future to look over complex social systems like misinformation online and determine how best to change policies to induce optimal outcomes.

REFERENCES

Alba, D. (2021, October 14). YouTube's stronger election misinformation policies had a spillover effect on Twitter and Facebook, researchers say. The New York Times. https://www.nytimes.com/2021/10/14/technology/distortions-youtubepolicies.html?searchResultPosition=2.

Au, C. H., Ho, K. K. W., & Chiu, D. K. W. (2021). Stopping healthcare misinformation: The effect of financial incentives and legislation. Health Policy, 125(5), 627–633. https://doi.org/10.1016/j.healthpol.2021.02.010

Chen, E., Chang, H., Rao, A., Lerman, K., Cowan, G., & Ferrara, E. (2021). Covid-19 misinformation and the 2020 U.S. presidential election. Harvard Kennedy School Misinformation Review. https://doi.org/10.37016/mr-2020-57

Crawford, T. H. (2020). Actor-network theory. Oxford Research Encyclopedia of Literature. https://doi.org/10.1093/acrefore/9780190201098.013.965

Fallis, D. (2015). What is disinformation? Library Trends, 63(3), 401–426. https://doi.org/10.1353/lib.2015.0014

Ito, J. (2011). Collective action for local commons management in rural Yunnan, China: Empirical evidence and hypotheses using evolutionary game theory. Land Economics, 88(1), 181–200. https://doi.org/10.3368/le.88.1.181

Jasper, J. (2004). A strategic approach to collective action: Looking for agency in social movement choices. Mobilization: An International Quarterly, 9(1), 1–16. https://doi.org/10.17813/maiq.9.1.m112677546p63361

Valenzuela, S., Halpern, D., Katz, J. E., & Miranda, J. P. (2019). The paradox of participation versus misinformation: Social media, political engagement, and the spread of

misinformation. Digital Journalism, 7(6), 802–823. https://doi.org/10.1080/21670811.2019.1623701

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media. ACM SIGKDD Explorations Newsletter, 21(2), 80–90. https://doi.org/10.1145/3373464.3373475