Undergraduate Thesis Prospectus

The Emergent Effects of Misinformation on Twitter

(sociotechnical research project)

by

Nicholas Tung

October 27, 2023

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

Nicholas Tung

STS advisor: Peter Norton, Department of Engineering and Society

General research problem

How can high-trust online spaces be developed?

Online communication mediums pose unique challenges in establishing and maintaining trusted discourse. Concurrently, online communication mediums enable connections between millions of users worldwide (Cheshire, 2011); increasing the baseline trustworthiness of online communication could have a broad impact on Internet users. Trustworthiness indicators in online spaces diverged from real-life indicators a long time ago; for example, domain name reputation has been a proxy for trustworthiness since the era of widespread Usenet usage and continues today (Donath, 1996). Studying mechanisms of establishing trustworthiness in online social contexts could mitigate deception and promote more valuable discourse.

Software for Project L social features

What lesson of lasting professional value did I learn from my internship experience at Riot Games?

The internship option is preferred over the CS curriculum option for the technical component of the prospectus. It isn't set in stone though.

The Emergent Effects of Misinformation on Twitter

How have social groups advanced their agendas through the dissemination of misinformation on Twitter (X)?

Compared to platforms with tighter-knit networks and/or decreased focus on discussion, Twitter's open nature creates a unique online space for discourse among disparate types of users. The platform has also played a pivotal role as a medium for information dissemination from public figures, news organizations, and citizens alike during elections (Jungherr, 2016). When groups with varying interests can directly interact with one another, how do they advance their own agendas surrounding information on Twitter (X)?

Since its founding in 2006, Twitter has propagated dangerous misinformation (Vosoughi et al., 2018). Researchers have also studied the way that Twitter's moderation policies have changed over time, showing how Twitter approaches moderation of misinformation with a lighter touch through downranking of tweets and addition of context labels (de Keulenaar et al., 2023). Finally, studies on how citizens interact with misinformation on Twitter (specifically, accusing another user's tweet of containing misinformation) has been explored and tied to an individual's locomotion orientation (Galande et al., 2023).

Twitter imposes its own misinformation ideas and values on its users through platform-wide systems, including employment of news organizations to write trustworthy context for trending topics (Geary, 2021) and "Community Notes", which combines user-submitted notes and ratings using a kind of bridging algorithm to surface context considered helpful by a broad set of users (Wojcik et al., 2022). To users, Twitter is both a communication channel and an entity that can take actions. Methods of affecting users aren't limited to the aforementioned proactive measures; policies like Musk-proclaimed algorithmic link punishment (Musk,) and the removal of link headlines (Peters, 2023) are hypothesized to change the flow of information on the platform.

Incentive structures also matter, especially in light of the platform's creator revenue sharing introduction. Released in August 2023, the program rewards creators for driving advertising revenue through their content (Lawler, 2023). While monetary incentives in social media contexts are not well studied, the effects of ad incentives are prominent in other content verticals such as network media and YouTube (Evans, 2019). The effects of revenue sharing on Twitter content may follow similar patterns.

Internet users relevant to the stated research question can be separated into two major classes: content creators and content consumers. The two are not mutually exclusive, but many Twitter users fall exclusively into the latter class. A 2019 study found that “most users rarely tweet, but the most prolific 10% create 80% of tweets from adult U.S. users” (Wojcik & Hughes, 2019). Members of these classes differ significantly in both behavior and agenda, as creators generally have significantly increased social or monetary capital to gain from their behavior compared to consumers. These two classes can be further divided into subgroups pertinent to misinformation on Twitter. Creator groups include active disinformation spreaders, traditional news organizations, and civilian informational creators. In the context of divisive topics, consumers can be separated by topical ideology. However, a generalizable distinction pertinent to Twitter’s information spread is between consumers in favor of platform moderation and consumers against platform moderation.

Disinformation spreaders may not be writing directly on Twitter; they could be publishing content elsewhere and using social media for distribution (Silverman & Alexander, 2016). They may be motivated by values pertaining to misinformation. NPR correspondent Laura Sydell interviewed Jestin Coler, founder and CEO of Disinfomedia. Coler said that his various fake news websites were intended to “‘infiltrate the echo chambers of the alt-right, publish blatantly or fictional stories and then be able to publicly denounce those stories and point out the fact that they were fiction’” (Sydell, 2016). Others indicate primarily monetary motivation: a Macedonian teenager running fake news website dailynewspolitic.com told BuzzFeed News that “‘in Macedonia the revenue from a small site is enough to afford many things'" (Silverman & Alexander, 2016). Both Coler and the unnamed Macedonian teenagers relied on inbound traffic from social media websites, primarily Facebook, to earn ad revenue from their fake news websites. These disinformation websites have an antagonistic relationship with media platforms, with, for example, Twitter actively blocking unapproved links.

Fake news websites differ significantly from traditional news organizations in content, but not so much with respect to operational structure. News organizations primarily run websites that serve ads or host paywalled journalism. Today, they are also heavily reliant on inbound traffic from algorithmic content delivery platforms like search engines and social media; 65% of page views that publishers received in 2023 came from search and social media (Majid, 2023). News organizations' historically tenuous relationship with algorithmic content platforms is visible through tech company initiatives like Facebook's pivot to video and Google's AMP project. Both events show how news organizations had their hands forced by tech companies in unfavorable ways. An article published by *The Atlantic* states that Facebook's inflation of video metrics led many media companies to make "the disastrous decision to 'pivot to video,'" which backfired when "views plunged and video's poor return on investment became more apparent" (Madrigal & Meyer, 2018). In Google's case, the purported search prioritization of Google's own website standard effectively forced media companies to adopt AMP while also limiting "'control over UI, monetization et al.,' said one digital media executive" (Ingram, 2016). This relationship holds true today, as Meta and Google platforms pull back from news following legislative movements to: punish platforms for spreading misinformation and require payment for publisher content (Fischer, 2023).

All this is to show that 2023 Twitter policy changes may similarly force the hands of news organizations that distribute content on the platform. It is no secret that Twitter owner Elon

Musk treats traditional news organizations with derision, calling NPR “state-affiliated” and referring generally to such organizations’ content as “legacy news” (Robison, 2023). When Twitter stopped showing titles for links to external websites for “aesthetic” purposes, journalism professor Karin Wahl-Jorgensen suggested that the change “can be seen as part of a larger trend toward making Twitter/X more difficult for news organizations to use” (Sands, 2023). *Slate* writer Alex Kirshner expressed concerns that the change “creates obvious opportunities for people to lie about or dramatize where a link will take them,” deceiving and confusing users (Kirshner, 2023). While Kirshner also states that this change is “a tiny deal” as “Twitter is famous in media circles for being a paltry source of web traffic,” industry professionals aren’t optimistic about its effects on the usability of Twitter as a platform for current events discourse (Kirshner, 2023).

Finally, traditional news organizations have fiduciary and integrity obligations; if nothing else, large news organizations built their brand and business on a reputation for quality journalism. This limits their actions and raises stakes, obstacles that the third relevant participant group, civilian informational creators, are not subject to. University of Washington’s Center for an Informed Public compared the reach of traditional news organizations with that of identified “highly influential accounts in the Hamas/Israel discourse on X,” finding that these non-news organization accounts generated greater engagement than traditional news accounts with over ten times the following. This significant source of information on a current event has unique characteristics: the report found that “the majority of the accounts and their tweets rarely included cited sources” while others mention sources “without giving any external links to them”. The content tended to be brief, punchy, and “emotionally charged” (CIP, 2023). These accounts appear to be driven by similar reporting style values. At the time of writing, every studied account is also a X Premium subscriber, granting a blue checkmark and eligibility for ad revenue sharing. This apparent incentive is blamed for X's transformation "into a hive of so-called engagement farming" (Lee et al., 2023). This style of content creation is not unique to the Hamas/Israel discourse, with similar accounts tweeting about tech and politics.

On the consumer side, most online media consumers note the misinformation issue and want solutions from platforms (St. Aubin & Liedke, 2023). In many cases, creators are also consumers; such creators have expressed concern with Twitter's lax protection from misinformation (Kirshner, 2023). *Mashable* writer Matt Binder collects a series of Twitter users' reactions to the recent link headline change that include disdain, complaints, and memes (Binder, 2023). On the other hand, Twitter's "Community Notes" feature has been met with mixed reactions: while many appreciate the intention, there is a sentiment that the feature, which only affects a minority of misinformation on Twitter, isn't enough (Goggin, 2023). One Community Notes volunteer expressed concerns that the two days it took for "the backroom to press whatever button to finally make all our warnings publicly viewable" was too long (Goggin, 2023). The algorithm determining helpfulness and whether or not a Community Note should be viewable is supposed to be fully-automated (Wojcik et al., 2022), but users are not fully satisfied.

Platform users against moderation efforts may be motivated by a diverse set of values or interests. For example, some argue that inaccurate classification of misinformation harms the platform as a whole, rendering such efforts more dangerous than no moderation (Lorenz, 2022). Others may have been subjects of intended moderation tools like account suspensions or bans. Aforementioned Twitter user @WarMonitors had their account limited on October 19th and "didn't provide a reason" beyond "you have violated the Twitter Rules" (War Monitors, 2023), and seemingly left the platform following a cryptic tweet on October 20th (War Monitors, 2023).

Deplatforming reduces their ability to influence others through their content (Jhaver et al., 2021). Finally, there are people who value a particular flavor of “free speech” and view moderation as an infringement of free speech rights; prominent members of this group include Elon Musk pre-Twitter acquisition (Robertson & Radtke, 2022).

References

Cheshire C. (2011). Online trust, trustworthiness, or assurance?. Daedalus, 140(4), 49–58. https://doi.org/10.1162/daed_a_00114

de Keulenaar, E., Magalhães, J. C., & Ganesh, B. (2023). Modulating moderation: a history of objectionability in Twitter moderation practices. *Journal of Communication*, *73*(3), 273-287.

Fischer, S. (2023, June 27). Social media news consumption slows globally. Axios. https://www.axios.com/2023/06/27/social-media-news-consumption-slows-globally

Galande, A. S., Mathmann, F., Ariza-Rojas, C., Torgler, B., & Garbas, J. (2023). You are lying! How misinformation accusations spread on Twitter. *Internet Research*.

Geary, J. (2021, August 2). Bringing more reliable context to conversations on Twitter. Twitter Blog. https://blog.twitter.com/en_us/topics/company/2021/bringing-more-reliable-context-to-conversations-on-twitter

Ingram, M., & Baumgarten, U. (2016, August 16). Google Says It Wants to Help Publishers, But Some Remain Skeptical. Fortune. https://fortune.com/2016/08/16/google-publishers-amp/

Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-30. https://dl.acm.org/doi/10.1145/3479525

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, *13*(1), 72-91. https://www.tandfonline.com/doi/full/10.1080/19331681.2015.1132401

Lawler, R. (2023, August 10). Elon Musk's new round of X Ads Revenue Sharing payments arrived — eventually. The Verge.

https://www.theverge.com/2023/8/11/23824612/x-twitter-blue-ad-revenue-sharing-payment-delay

Lee, D., Kaiser, A. J., Molero, G., & Chua, H. (2023, October 5). The Moral Case for No Longer Engaging With Elon Musk's X. Bloomberg.com. https://www.bloomberg.com/opinion/articles/2023-10-05/the-moral-case-for-no-longer-engaging-with-elon-musk-s-x#xj4y7vzkg

Lorenz, T. (2022, August 25). Twitter labeled factual information about covid-19 as misinformation. *Washington Post*. https://www.washingtonpost.com/technology/2022/08/25/twitter-factual-covid-info-labeled-misinformation/

Madrigal, A. C., & Meyer, R. (2018, October 18). The Facebook-Driven Video Push May Have Cost 483 Journalists Their Jobs. The Atlantic. https://www.theatlantic.com/technology/archive/2018/10/facebook-driven-video-push-may-have-cost-483-journalists-their-jobs/573403/

Majid, A. (2023, April 13). How do consumers find news? News referral traffic breakdown. Press Gazette. https://pressgazette.co.uk/media-audience-and-business-data/media_metrics/news-referral-traffic-breakdown/

Peters, J. (2023, October 4). X stops showing headlines because Elon Musk thinks it will make posts look better. The Verge. https://www.theverge.com/2023/10/4/23903859/x-elon-musk-headlines-links-image-twitter

Robertson, A., & Radtke, K. (2022, April 15). What Elon Musk's Twitter 'free speech' promises miss. *The Verge*. https://www.theverge.com/2022/4/15/23025120/elon-musk-twitter-free-speech-government-censorship

Robison, K. (2023, October 6). Elon Musk's secret PR machine at X. Fortune. https://fortune.com/2023/10/06/elon-musks-secret-pr-machine-x-twitter/

Sands, L. (2023, October 5). Elon Musk removes news headlines from displaying on X, formerly Twitter. The Washington Post. https://www.washingtonpost.com/technology/2023/10/05/twitter-x-news-headlines-removed/

Silverman, C., & Alexander, L. (2016, November 3). How Teens In The Balkans Are Duping Trump Supporters With Fake News. *BuzzFeed News*. https://www.buzzfeednews.com/article/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo#.nfGBdzv3rN

St. Aubin, C., & Liedke, J. (2023, July 20). *Most favor restricting false information, violent content online*. Pew Research Center. https://www.pewresearch.org/short-reads/2023/07/20/most-americans-favor-restrictions-on-false-information-violent-content-online/

Sydell, L. (2016, November 23). We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned. *NPR*. https://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs

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

War Monitor [@WarfareBackup]. (2023, October 19). *Twitter just locked me for 12 hours and didn't provide a reason* [Image attached] [Tweet]. Twitter. https://twitter.com/WarfareBackup/status/1715078469300744240

War Monitor [@WarMonitors]. (2023, October 20). *Till next time.* [Image attached] [Tweet]. Twitter. https://twitter.com/WarMonitors/status/1715344721164349846

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Fallin Munzaker, M.B., Coleman, K., & Baxter, J. (2022). Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation. arXiv. https://doi.org/10.48550/arXiv.2210.15723

Wojcik, S., & Hughes, A. (2019, April 24). How Twitter Users Compare to the General Public. Pew Research Center. Retrieved October 28, 2023, from https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/