How Social Groups Spread Misinformation on Twitter

An STS Research Paper

Presented to the faculty of the

School of Engineering and Applied Science

University of Virginia

By

Nicholas Tung

March 14, 2024

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

Nicholas Tung

STS Advisor: Peter Norton

Introduction

Online communication mediums pose unique challenges in establishing and maintaining high-trust spaces and communication. In computational trust, privacy, identity, and permission can promote trustworthiness. Computational trust generally cannot be guaranteed given a large population of diverse technological literacy; for example, nonexperts may use end-to-end encryption in popular messaging applications ineffectively (Dechand et al., 2019). By contrast, generalized trust can more readily be established online.

Schilke et al. (2021) defined trust as "the willingness of an entity (i.e., the trustor) to become vulnerable to another entity (i.e., the trustee)." More specifically, generalized trust "typically involves a relatively large circle of unfamiliar others." Trust can readily develop in small communities or organizations. A version of such community trust can emerge on social networks. A user's social network graph could be considered a "large circle of unfamiliar others," especially after considering second and third degree connections that may be surfaced by a platform. Nevertheless online trust is fragile. Because social media supports connections between millions of users worldwide (Cheshire, 2011), increasing the baseline trustworthiness of online communications could have vast implications.

How can users' trust in online entities be increased? Research suggests that trust between entities is influenced by a variety of factors categorized into the following groups: trustor factors, trustee factors, and contextual factors (Hancock et al., 2023). At first glance, the trustor-trustee relationship seems difficult to analyze generally. However, looking at trust with a contextual focus reveals idiosyncrasies of online interaction compared to traditional interaction patterns. Social graphs tend to be wider online, increasing connection quantity but limiting depth. Additionally, there is limited ability to use social cues to gauge trustworthiness. 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). To build such trust online, the affordances of the specific medium must be considered as part of the greater task. Whether through culture or enforcement, studying mechanisms of establishing trustworthiness in online social contexts could mitigate deception and promote more valuable discourse.

There are various platforms that fall under the umbrella of social networks. There are some commonalities between the modern-day platforms, but each provides its own affordances. With the importance of specific platform affordances in mind, the scope of this paper is narrowed further to focus on the platform known as Twitter (renamed 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: famous or unknown, wealthy or poor, people from just about any background can access Twitter with a computer and Internet access. The platform's other provisions incentivise interesting behavior in its users; an illuminating example is the ability to quote other users' tweets, which fragments the location of discussions, elevates visibility of controversial tweets, and often leads to users "dunking" on others (McNear, 2018). Its affordances, impact, and incentives make it a noteworthy platform for further study.

Observing Internet discourse through the lens of a platform's affordances, limitations, and cultures provides interesting insight while limiting scope. In Twitter's case, however, this lens is still too wide for this paper. Twitter provides many functions. For a particular user, Twitter could provide any subset of news, discourse, community, comedy, education, and opinion. Anyone can talk to anyone on Twitter, making the platform relatively open, but the emergent network of

voices is different for each user. Some Twitter content could be seen as a window into the zeitgeist of the modern era, while other categories of content could be viewed as fringe and isolated. Twitter's breadth necessitates further focus on one function: information dissemination. The platform has played a pivotal role as a medium for information dissemination from public figures, news organizations, and citizens alike during elections (Jungherr, 2016). Unlike niche communities or humor, information and understanding of the world is within the purview of many users.

Twitter plays a role in news and information dissemination but also allows just about anyone to make an account and start tweeting. By placing tweets from news organizations and public figures alongside tweets from ordinary citizens in a minimally filtered feed, Twitter creates a potent environment for misinformation to spread. Vectors for this spread include accounts impersonating or mimicking legitimate sources as well as apparent individuals simply making assertions. The importance of trusted communication on Twitter is clear, but properties of Twitter make trusted communication and information dissemination non-trivial. Why hasn't Twitter changed to increase trusted communication and minimize misinformation? Many would agree that a well-informed populace is a good thing, so where are the solutions from interest groups like the platform and the government? Twitter doesn't want misinformation (*Our Synthetic and Manipulated Media Policy | X Help*, 2023), but it proliferates regardless. Understanding why misinformation continues to spread requires studying Twitter's incentive structures. Twitter rewards particular behavior from various interest groups and limits the faculties of opposing groups in ways that enable the continued dissemination of misinformation.

Review of Research

At its core, this paper is examining communication mechanics with respect to a particular platform and category. Media studies and general communication research is relevant, despite possible chronological distance between the research and the current day. Consider the one-step and two-step communication flow models. Both models were conceived in the mid-20th century, well before the advent of algorithmic content recommendation. The one-step flow supposes that media and its messages travel from the source directly to its audience. The two-step flow differs from the aforementioned model, however, and supposes that messages reach the general audience through opinion leaders rather than mass media. These models aren't mentioned because one or the other has proven accurate. They are, however, useful frameworks for thinking about the flow of information from a source to an audience. Bennett et al. states that in the United States, the two-step flow was more accurate during the era of network news and mass media, while the one-step flow became increasingly relevant during the advent of algorithmic content recommendation platforms.

Increasing specificity from media studies research, research on Twitter is unsurprisingly relevant. Twitter is a well-studied platform; its cultural relevance and past ease of access via its application programming interface (API) make it a pragmatic target for research. Study of over 150,000 tweets related to environmental social movements in Chile revealed that while a significant portion of platform communication follows the two-step flow, 20% of the total communication occurs through direct channels in a one-step flow (Hilbert et al., 2017). On Twitter, other adjacent communication models emerge from the data, indicating that different models of communication exist alongside each other, with the communication flow in a particular instance "depending on perspective" (Hilbert et al. 2017).

Research has been conducted on misinformation in particular on Twitter. 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) have been explored and tied to an individual's locomotion orientation (Galande et al., 2023).

Finally, Twitter is not the only algorithmic content recommendation platform. While mediums like YouTube and Facebook have different affordances and limitations, some properties and effects related to those platforms are relevant to this discussion of Twitter. In August 2023, Twitter unveiled its creator revenue sharing program, which 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. With respect to misinformation, Youtube, Facebook, and other platforms are also vectors of misinformation dissemination. Like Twitter, these platforms enable wide swaths of users to publish content with minimal monitoring or filtering (Gant, 2007).

Creators have few incentives for trustworthiness

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. An individual following can be monetized in a variety of ways; such mechanisms form the basis of the creator economy. This over 100 billion dollar market relies on content that is perceived as individual and authentic, compared to more traditional media content (Peres et al., 2023). The creator class can be further divided into subgroups pertinent to misinformation on Twitter. Creator groups include active disinformation spreaders, traditional news organizations, and civilian informational creators.

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). For disinformation content motivated by such an ideology, Twitter is more a medium than an incentivising force. By providing a platform that not only allows citizens like Coler to hypothetically tweet arbitrary links but also houses alt-right echo chambers, Twitter enables this particular interaction with minimal friction, as there is no incentive against such behavior.

Other disinformation spreaders 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. Twitter, with its light-handed moderation policies (de Keulenaar et al., 2023), pays relatively little attention to accuracy of small websites and provides inbound traffic for sites like dailynewspolitic.com.

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. 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). History shows how news organizations must follow directives set by platforms or face the consequences.

All this is to show that 2023 Twitter policy changes may similarly force the hands of news organizations that distribute content on the platform. Twitter policy and attitude toward news organizations handicaps these media producers in ways that do not affect other types of creators the same way. 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" (Caulfield et al., 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.

Minimal consumer protections

Twitter ultimately allows and even incentivises creators to peddle misinformation. However, submitting content to algorithmic recommendation platforms does not imply that the content receives an audience. Platforms also do not guarantee that content will be shown to an audience unmodified. This is standard behavior for modern platforms: YouTube regulates misleading content and has taken down content in the past, and Meta similarly takes down and fact checks content (Swenson & Fernando, 2023). Twitter imposes its own misinformation ideas and values on its users through platform-wide systems. Like Meta, Twitter fact checks high-visibility trending topics by employing news organizations to write trustworthy context (Geary, 2021). Unfortunately, topics trend because of significant engagement with a class of content, meaning that fact checking topics once they are trending has limitations.

Fact checking trending topics is a top-down approach to fighting misinformation. In the past few years, a different Twitter feature has been gaining prominence: Community Notes. Formerly known as Birdwatch, Community Notes enables users, rather than just employees, to submit context for tweets. These context submissions are then rated by other Twitter users for

helpfulness. A unique facet of Community Notes is the algorithm used to surface notes based on ratings. Rather than a naive voting system (helpful or unhelpful), Community Notes uses a kind of bridging algorithm that optimizes for notes that are rated helpful by a broad set of users. Specifically, this algorithm effectively reduces the rating of polarizing notes that are rated helpful by one class of Community Notes raters but not a different class of raters (Wojcik et al., 2022). When executed as described in the paper, Community Notes surfaces objective and helpful context. Since Community Notes is also an automated system, it scales better than features that rely on Twitter employees or third-party organizations.

While innovative, Community Notes is far from perfect. Similar to fact checking of trending topics, Community Notes take time to be shown on content. During the time that misinformation is visible without a Community Note, it is still spread among users. 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). A typical note correcting misinformation regarding the Israel-Hamas conflict took "more than seven hours to show up, while some took as long as 70 hours" (Alba et al., 2023). In that sense, Community Notes can slow the spread of misinformation but is structurally incapable of preventing its spread before publication.

The requirements for a Community Note to be shown alongside a tweet hamper the program's effectiveness in more ways than one. Beyond taking hours to show a note, many submitted Community Notes are never shown to the general populace in the first place. Of the tens of thousands of notes written by contributors, approximately 96% of them remain hidden (Fan et al., 2022). There are bound to be submitted notes that do not meet the intended quality bar, but the broad consensus requirements for Community Notes means that many useful notes

are also never surfaced. Notes that cite sources and contest claims made in a tweet require raters across the political spectrum to rate them helpful. Fan et al. show an example note that corrects a tweet that states “Abortion is never medically necessary to save the life of a mother”. Data published by Twitter shows that 97% of raters rated the note helpful, but the note was not shown because the vast majority of those raters had similar viewpoints. Preventing notes that have cross-ideological consensus from being shown limits abuse, but it is a double-edged sword that hides objectively helpful notes.

The broad consensus requirements reveal another limitation of Community Notes: it often fails to address the most divisive tweets. While rater helpfulness (a measure of how often a particular user’s ratings matched the consensus in the past) is considered in determining a note’s rating (Wojcik et al., 2022), since the program operates on volunteer ratings, users can choose not to rate notes for whatever reason and prevent it from becoming visible (Fan et al., 2022). Fan et al. found that Community Notes are typically only surfaced alongside “comparatively mild” content. This doesn’t make Community Notes a failure, but it shows how the scope of Community Notes is limited to a particular class of misinformation. With Community Notes becoming a more prominent feature of Twitter, the absence of a note carries meaning. It is possible that users are more likely to accept a tweet as truth because it lacks a Community Note, treating the omission as an endorsement of sorts. Assessing the impact (or lack thereof) of Community Notes on misinformation requires continued study that considers its nuances.

Conclusion

With everything in mind, it isn’t surprising that misinformation continues to spread on Twitter. Rather than being merely an issue caused by some set of malicious Twitter users,

structural facets of the platform enable the proliferation of misinformation. Users are incentivised in various ways to tweet misinformation, and sources of correct information (like news organizations) face unique limitations that affect their reach. Misinformation content is thus bound to be submitted to the platform, and Twitter doesn't do enough to protect consumers from it. People are bound to see uncorrected misinformation, with important context appearing after a delay if at all.

References

Alba, D., Lu, D., & Yin, L. (2023, November 20). How Community Notes on Twitter Are Failing to Combat Israel-Hamas War Misinformation. *Bloomberg.com*. https://www.bloomberg.com/graphics/2023-israel-hamas-war-misinformation-twitter-community-notes/

Bennett, W. L., & Manheim, J. B. (2006). The One-Step Flow of Communication. The ANNALS of the American Academy of Political and Social Science, 608(1), 213-232. https://doi.org/10.1177/0002716206292266

Caulfield, M., Bayar, M. C., & Aske, A. B. (2023, October 20). *The 'new elites' of X: Identifying the most influential accounts engaged in Hamas/Israel discourse*. UW Center for an Informed Public. Retrieved March 14, 2024, from https://www.cip.uw.edu/2023/10/20/new-elites-twitter-x-most-influential-accounts-hamas-israel/

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.

Dechand, S., Naiakshina, A., Danilova, A., & Smith, M. (2019). In Encryption We Don't Trust: The Effect of End-to-End Encryption to the Masses on User Perception. 2019 IEEE European Symposium on Security and Privacy (EuroS&P), 401-415. 10.1109/EuroSP.2019.00037

Fan, E., Dottle, R., & Wagner, K. (2022, December 19). Twitter's Fact-Checking System Has a Major Blind Spot: Anything Divisive. *Bloomberg.com.*

https://www.bloomberg.com/graphics/2022-twitter-birdwatch-community-notes-misinformation-politics/

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*.

Gant, S. (2007). *We're all journalists now: the transformation of the press and reshaping of the law in the internet age.* Simon and Schuster.

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

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Stowers, K., Brill, J. C., Billings, D. R., Schaefer, K. E., & Szalma, J. L. (2023). How and why humans trust: A meta-analysis and elaborated model. Frontiers in psychology, 14, 1081086. https://doi.org/10.3389/fpsyg.2023.1081086

Hilbert, M., Vásquez, J., Halpern, D., Valenzuela, S., & Arriagada, E. (2017). One Step, Two Step, Network Step? Complementary Perspectives on Communication Flows in Twittered Citizen Protests. *Social Science Computer Review,* 35(4), 444-461. http://dx.doi.org/10.1177/0894439316639561

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/

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

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/

McNear, C. (2018, May 2). How Quote Tweets Helped Ruin Twitter. The Ringer. https://www.theringer.com/tech/2018/5/2/17311616/twitter-retweet-quote-endorsement-function-trolls

Our synthetic and manipulated media policy | X Help. (2023, April). Twitter Help Center. Retrieved March 15, 2024, from https://help.twitter.com/en/rules-and-policies/manipulated-media

Peres, R., Schreier, M., Schweidel, D. A., & Sorescu, A. (n.d.). The Creator Economy: An Introduction and a Call for Scholarly Research. SSRN Electronic Journal. 10.2139/ssrn.4663506

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

Swenson, A., & Fernando, C. (2023, December 26). Misinformation may get worse in 2024 election as safeguards erode. *AP News*. https://apnews.com/article/election-2024-misinformation-ai-social-media-trump-6119ee6f498db10603b3664e9ad3e87e

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

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