Surfacing Trust, Assessment, and Provenance for Better News Sharing

Farnaz Jahanbakhsh
farnazj@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Amy X. Zhang
axz@cs.uw.edu
University of Washington
Seattle, WA, USA

David R. Karger
karger@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

ABSTRACT

To employ users as autonomous allies of platforms in their battle against misinformation, we consider three design choices that could be integrated into social media and news platforms to give users more power and agency: capturing users' assessments of posts as part of the data model, allowing them to explicitly indicate who they trust to assess sensibly, and user-operated manual filtering by accuracy assessments into the platforms' data models and user interfaces. We evaluate these design changes via a need-finding study of more than 150 users and a user study of 14 participants on a platform that incorporates these design decisions. We then discuss the challenges and potentials of our design changes.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

KEYWORDS

Misinformation, Social Media, Fact-checking, News Reading and Sharing Platform

ACM Reference Format:
Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Surfacing Trust, Assessment, and Provenance for Better News Sharing. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 USERS AS AUTONOMOUS ALLIES OF PLATFORMS

As social media platforms track down misinformation using ML models and human fact-checkers [2, 3], they forgo the help that they could be receiving from one of the primary stakeholders in the space: the users. Indeed, the engagement-based model of platforms even undermines users' efforts to combat misinformation. For instance, a user who comments on a post to dispute its veracity can be unknowingly spreading the post further because the system considers the interaction as engagement.

In this work, we attempt to re-imagine the platforms used for news sharing and consumption by giving users more agency to protect themselves and their social circle against misinformation. We consider three complementary design changes that could be incorporated into platforms:

- allowing users to explicitly assess posts as accurate or inaccurate as part of the data model
- enabling them to explicitly mark other sources—users or news publishing entities— as trustworthy
- providing them with filters which they can use to filter out content assessed as inaccurate by their trusted sources from their feed

Integrating trust assessments of posts into the data model can help disambiguate not only engagement vs refutation of a post for the platforms, but also the various signals that could be mistaken as one or the other by the other users. For instance, a "laugh" emoji on a post could signify disdain at an incorrect post or conversely, approval of its remarks. However, an "inaccurate" tag can clearly make the distinction. Asking users to assess posts, in addition to signaling post accuracy, can help by curbing the propagation of falsehood altogether. In an experiment, we evaluated the effects of a set of behavioral nudges that could be incorporated in social media upon sharing time on people's intention of sharing a post. We found that asking users to assess accuracy of news items before sharing them reduces their likelihood of sharing misinformation. Further asking people for their rationales decreases their sharing of falsehood even more [10]. This finding has been corroborated in other studies that report nudging users to think about accuracy can help increase quality of news that they share [14, 16].

Our design choice to allow people to specify their trusted sources and receive fact-checking information from them is inspired by prior work that reports social media users are more likely to attend to and accept fact-checking information from friends compared to strangers [8, 13]. Indeed, correcting information from various other sources such as journalists often struggle to keep up with the misinformation that has spread [17].

The filters that we envision enable users to narrow their feed down to articles with a certain accuracy status, for instance, verified or refuted, assessed by the sources that they explicitly mark as trusted. This design of giving users agency over their feed is informed by prior work that reports users are more satisfied when given controls over their feed and that in the absence of control, try to find ways around the curation algorithm that chooses content for them [6, 18].

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 EVALUATION

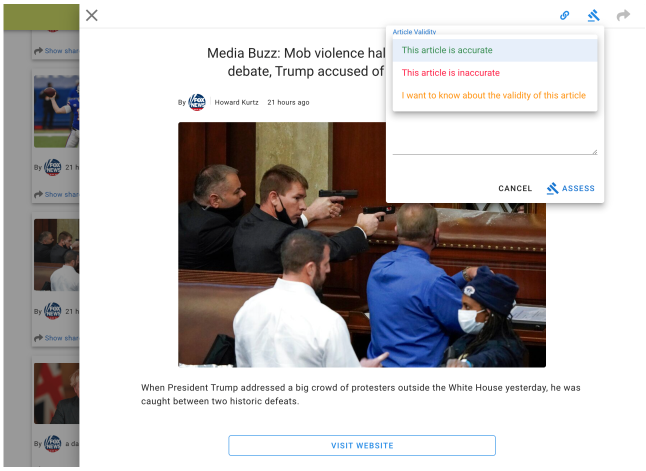
To evaluate the three design ideas described above, we performed a need-finding study in which we surveyed a diverse group of 157 users about their practices reading and sharing online news. We found that users' exposure to misinformation results from their intentionally following unreliable sources for a variety of reasons or because their otherwise reliable sources may make occasional mistakes. Some users would like to take the misinformation out of their feed but continue following these sources, and others would like to keep the unreliable content in their feed along with signals of the content's inaccuracies. These wishes suggest the need for filters that empower users to take more control over their feed. In addition, users already engage in soliciting fact-checking information from as well as providing it to their social circle. Some expect their social network to proactively correct them should they post inaccuracies. However, because social media platforms do not have designated metadata for assessing accuracy of posts, participants use a variety of features intended for other purposes, such as likes and comments, that can be picked up by the platforms as signals of engagement.

We then conducted a technology probe into the problem space by conducting a user study on a platform we built that embraces the paradigms described above, where users can mark posts as accurate or inaccurate or inquire about the validity of posts, specify the sources they trust, and filter their feed based on accuracy assessments of posts provided by their trusted sources.

The platform treats users and proxy accounts that it manages on behalf of news publishing entities (e.g., the NYTimes) as sources of news. It provides two types of relationships between sources: follow and trust. By following another source, a user can see what the source posts. Trusted sources however, can be leveraged for filtering one's newsfeed based on the accuracy assessments that they provide. The two relationships are asymmetric and independent from each other. Therefore, a source can follow but not trust another source, or can trust but not follow them. Additionally, a source that trusts another is not necessarily trusted by the latter source. Trust relationships on our platform are kept private which can ease the social pressure of users marking someone they do not trust as trustworthy. Public trust relationships on the other hand, can be beneficial for public endorsements, for instance between journalists, researchers, or news publishing entities, and can be explored in future work. Users can additionally group sources into lists that are private to them.

On the platform, users can write posts of their own or import articles from other websites. In addition, the platform allows users to add RSS feeds into the system. It then makes a source associated with the added feed and periodically fetches the contents of the feed based on an estimate of how frequently the feed updates its contents.

All sources on the platform can assess posts as accurate or inaccurate. They can additionally inquire about the validity of a post. This inquiry can be anonymous or with the source's name attached to it. Requests for assessments are by default surfaced to the source's trusted sources. However, the source making the inquiry can specify from which sources they would like to receive assessments, in which case the specified sources would be notified of the question. No source can share a post before assessing it. A source's profile contains all the posts they have ever assessed and shared. Figure 1 shows the UI for assessing a post.



Media Buzz: Mob violence hal
debate, Trump accused of

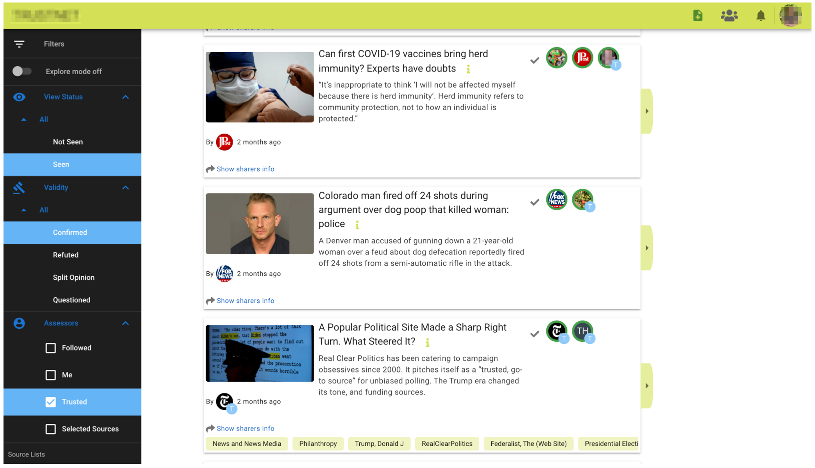
By Howard Kurtz 21 hours ago

When President Trump addressed a big crowd of protesters outside the White House yesterday, he was caught between two historic defeats.

Figure 1: The UI for assessing a post. When asserting that the post is accurate or inaccurate, a rationale is required. Assessing is required before sharing is enabled.

On their homepage, users see content from the sources that they follow. The platform provides filters that they can use to narrow down the articles on their feed. Users can filter posts based on their validity—assessed as accurate, inaccurate, a mixed of accurate and inaccurate (split opinion), and those whose validity have been questioned (questioned), who has assessed them (trusted sources, followed sources, specific source lists or sets of sources, or the user themselves), whether the user has previously seen the posts, and the tags associated with the posts. Figure 2 shows the homepage with the filters in the sidebar. Figure 3 displays the expanded assessments pane for an article which contains the assessments given by the sources the user follows or trusts.

To recruit participants for the user study on the platform, we asked that participants join the study with at least one other member from their social circle. This was because we were interested in understanding how users interact with their social circle to provide and receive fact-checking information. Although users can be exposed to misleading or false information, accurate information from a source with a viewpoint opposing theirs, or inaccurate information from sources with similar viewpoints as theirs on social media platforms where their extended social circle is present, it was possible that without an intervention from our part, at such a small scale, participants would not be exposed to these types of content. Therefore, we increased their likelihood of exposure to such information by asking them to follow a source that we named Trending News which was managed by a member of our research team. Every day for the duration of the user study, the Trending News source imported and shared a number of articles from different sources with varying degrees of credibility. Participants' tasks for the duration of the user study (one week) involved using the platform to read and share at least two posts every day and checking whether anyone else has asked for their assessment. The UI



Can first COVID-19 vaccines bring herd immunity? Experts have doubts

"It's inappropriate to think 'I will not be affected myself because there is herd immunity'. Herd immunity refers to community protection, not to how an individual is protected."

By 2 months ago

Colorado man fired off 24 shots during argument over dog poop that killed woman: police

A Denver man accused of gunning down a 21-year-old woman over a feud about dog defecation reportedly fired off 24 shots from a semi-automatic rifle in the attack.

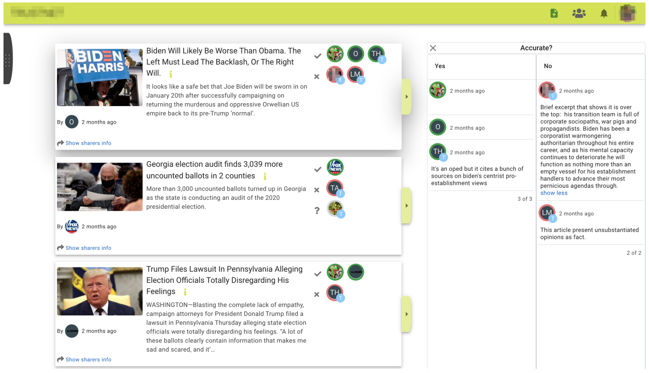
By 2 months ago

A Popular Political Site Made a Sharp Right Turn. What Steered It?

Real Clear Politics has been catering to campaign obsessives since 2000. It pitches itself as a "trusted, go-to source" for unbiased polling. The Trump era changed its tone, and funding sources.

By 2 months ago

Figure 2: The homepage view with articles filtered according to the filters on the left sidebar. Articles can additionally be filtered using tags (e.g., those on the New York Time's article in this screenshot).



Biden Will Likely Be Worse Than Obama. The Left Must Lead The Backlash, Or The Right Will.

It looks like a safe bet that Joe Biden will be sworn in on January 20th after successfully campaigning on returning the murderous and oppressive Orwellian US empire back to its pre-Trump 'normal'.

By 2 months ago

Georgia election audit finds 3,039 more uncounted ballots in 2 counties

More than 3,000 uncounted ballots turned up in Georgia as the state is conducting an audit of the 2020 presidential election.

By 2 months ago

Trump Files Lawsuit In Pennsylvania Alleging Election Officials Totally Disregarding His Feelings

WASHINGTON—Blasting the complete lack of empathy, campaign attorneys for President Donald Trump filed a lawsuit in Pennsylvania Thursday alleging state election officials were totally disregarding his feelings. "A lot of these ballots clearly contain information that makes me sad and scared, and it'...

By 2 months ago

2 months ago

It's an oped but it cites a bunch of sources on biden's centrist pro-establishment views

Brief excerpt that shows it is over the top: his transition team is full of corporate sociopaths, war pigs and propagandists. Biden has been a corporatist warmongering authoritarian throughout his entire career, and as his mental capacity continues to deteriorate he will function as nothing more than an empty vessel for his establishment handlers to advance their most pernicious agendas through.

This article present unsubstantiated opinions as fact.

Figure 3: An article tile shows a preview of which of the user's followed or trusted sources has assessed the article as accurate, inaccurate, and who has inquired about its validity. The assessment pane can be expanded to show their full assessments.

would remind the participants about their daily goals every few hours at random times if they had not yet achieved them.

A total of 14 participants completed our user study (8 males, 6 females). One identified as Republican, two as Independent, and the rest were Democratic. 7 other participants registered but dropped out either at the onset or in the midst of the study. Their demographics were similar to those who completed the study. Users completed a questionnaire at the end of the user study.

The user study revealed that users indeed saw value in the norms that the platform established such as vetting posts before sharing, in facilitating inquiries about validity of posts, and in seeing assessments from others. We observed that although we did not train participants in how to assess posts, they indeed used the assessments to evaluate the veracity of posts based on the rationales reported in [10]. In addition, the various preferences for setting the filters and the absence of a panacea —e.g., viewing all articles by default but occasionally selecting to see the disputed ones vs choosing to view only the confirmed articles by trusted sources by default and occasionally seeking unverified posts— indicated the need for putting users more in control of their feed.

The study also highlighted areas that could be improved by leveraging a combination of ML models and user input, for instance, for extracting claims from articles that contain multiple claims, and for labeling certain articles as satire or oped that perhaps cannot be assessed as factually correct or incorrect. One point of confusion about assessments leading to inconsistencies that our participants cited was that sometimes the headlines of articles did not match their content. In these cases, some participants based their assessments on the content and others on the headline.

3 DISCUSSION AND FUTURE WORK

Many proposed approaches to countering the misinformation problem on platforms involve identifying and removing content that are deemed by the platforms as unfit to be seen by their users. Although policy-driven platform moderation is necessary in some universally agreed upon contexts [15], communities should be wary of relinquishing all the power of content filtering and highlighting to the platforms whose incentives as for-profit entities running on ads do not necessarily align with the users' [7]. The challenge of moderation is exacerbated as not all accounts of problematic behavior or posts can be provisioned a priori in platform policies, leading moderators to make ad-hoc decisions in grey areas that sometimes draw criticism [1, 9]. In addition, while platform moderators and fact-checkers can play a valuable role in flagging content that has already spread and become visible, other measures are needed to restrain sharing of misinformation as it is being handed from user to user. These challenges suggest that the problem of misinformation could additionally be tackled at the user level.

While it is difficult to persuade users to migrate from their current news readers and monopolized social media to a new tool to consume and share news, the credibility cues and assessments implemented into the tool can perhaps best be leveraged if offered via a browser extension as users browse various existing news websites and social media platforms. An extension would also allow users to assess or see assessments on not only posts existing in the platform, but also those that they encounter on social media, helping us gain insight into how these assessments would be used in the wild through a field study. We are currently working on developing such an extension.

The currently incorporated manual filtering of articles in the platform is a design decision that is independent of trust and assessment. Assessments incorporated into current platforms could be used as input for their algorithmic feeds, without introducing manual controls. Future work can study how users perceive algorithmic feeds of this nature.

Empowering users to block information they do not trust could conceivably lead to stronger "filter bubbles" in which a closed group of like-minded users mutually reinforce each others' perspectives while their trust filters shield them from any divergent views. However, we observed in our user study that users do seek information that has been questioned or refuted, or shared by untrusted sources. Our design does not prevent them from doing so; it simply puts them more in control. They can decide when they wish to remain safe in their bubble and when they wish to explore. This control and the safe space of a community of like-minded people that users can enter and exit whenever they wish, or *epistemic respite*, can help prepare users to re-engage with differing views [4]. In an attempt to facilitate this re-engagement, future work can design and implement features aimed specifically at puncturing filter bubbles. Consider a situation where the majority of a user's friends have assessed an article as true, and one as false. In today's networks, that outlier friend's disagreement would likely be invisible in a sea of one-sided comments. However, if structured assessments were captured, then a system could surface the fact that the article is disputed, and could highlight the outlier assessment and the fact of its being in disagreement with the rest.

Prior work has examined transitivity of trust in social networks, for instance in the context of recommender systems or chat moderation [5, 11, 12]. We are planning to extend the concept of trusted sources on the platform to build a trust network for each user. In this scenario, when a user leaves a credibility assessment on a post, the platform can propagate that information to all the users that either immediately trust or have an indirect trust path to the assessor, while maintaining the assessor's anonymity, as a user's trust relationships on our platform are private to the user. This chain of trust could help users benefit from more extensive assessments even though they may not immediately know the benefactors. Transitive trust relationships will allow us to investigate several interesting research questions. One is how fast trust decays as the distance of two sources in the network who are connected by an implicitly inferred trust relationship increases. Another is how assessments from different sources, some not immediately connected to the user, should be weighted, aggregated, and presented to the user in an interprettable manner. One scenario could be that each trusted source is given an equal weight in deciding the accuracy of an article. Conversely, the user could decide that a particular source or rationale be given priority over others.

4 CONCLUSION

In summary, the contribution of this work lies in re-imagining the platforms used for news sharing and consumption, by involving users in the fight against misinformation and giving them agency to protect their social circle. Our approach involves designing and building systems that facilitate a technology probe into the problem space, so that we can understand user needs and observe their behaviors in a real setting.

REFERENCES

[1] [n.d.]. *Facebook apologises for blocking Prager University's videos.* https://www.bbc.com/news/technology-45247302

[2] [n.d.]. *How Our Fact-Checking Program Works.* https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works

[3] Mike Ananny. 2018. The partnership press: Lessons for platform-publisher collaborations as Facebook and news outlets team to fight misinformation. (2018).

[4] Natalie Ashton. [n.d.]. *Why Twitter is (Epistemically) Better Than Facebook.* https://www.logically.ai/articles/why-twitter-is-epistemically-better-than-facebook

[5] Alexander Cobleigh. 2020. TrustNet: Trust-based Moderation Using Distributed Chat Systems for Transitive Trust Propagation. (2020).

[6] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems.* 2371–2382.

[7] Jennifer Grygiel and Nina Brown. 2019. Are social media companies motivated to be good corporate citizens? Examination of the connection between corporate social responsibility and social media safety. *Telecommunications Policy* 43, 5 (2019), 445–460.

[8] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. 2014. Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations.. In *ICWSM.*

[9] Yasmin Ibrahim. 2017. Facebook and the Napalm Girl: reframing the iconic as pornographic. *Social Media+ Society* 3, 4 (2017), 2056305117743140.

[10] Farnaz Jahanbakhsh, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. 2021. Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media. In *To appear in proceedings of the 2021 CSCW Conference on Computer-Supported Cooperative Work and Social Computing.*

[11] Audun Jøsang, Elizabeth Gray, and Michael Kinateder. 2003. Analysing topologies of transitive trust. In *Proceedings of the First International Workshop on Formal Aspects in Security & Trust (FAST2003).* Pisa, Italy, 9–22.

[12] Guanfeng Liu, Yan Wang, and Mehmet Orgun. 2011. Trust transitivity in complex social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 25.

[13] Drew B Margolin, Aniko Hannak, and Ingmar Weber. 2018. Political fact-checking on Twitter: When do corrections have an effect? *Political Communication* 35, 2 (2018), 196–219.

[14] Mohsen Mosleh, Gordon Pennycook, Antonio A Arechar, and David G Rand. 2021. Cognitive reflection correlates with behavior on Twitter. *Nature communications* 12, 1 (2021), 1–10.

[15] Kari Paul. [n.d.]. *Pornhub removes millions of videos after investigation finds child abuse content.* https://www.theguardian.com/technology/2020/dec/14/pornhub-purge-removes-unverified-videos-investigation-child-abuse

[16] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.

[17] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 Proceedings* (2014).

[18] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–13.