Tweet Trajectory and AMPS-based Contextual Cues can Help Users Identify Misinformation

HIMANSHU ZADE, Human-Centered Design and Engineering (HCDE), University of Washington, USA
MEGAN WOODRUFF, HCDE, University of Washington, USA
ERIKA JOHNSON, HCDE, University of Washington, USA
MARIAH STANLEY, HCDE, University of Washington, USA
ZHENNAN ZHOU, School of Computer Science, University of Washington, USA
MINH TU HUYNH, HCDE, University of Washington, USA
ALISSA ELIZABETH ACHESON, HCDE, University of Washington, USA
GARY HSIEH, HCDE, University of Washington, USA
KATE STARBIRD, HCDE, University of Washington, USA

Well-intentioned users sometimes enable the spread of misinformation due to limited context about where the information originated and/or why it is spreading. Building upon recommendations based on prior research about tackling misinformation, we explore the potential to support media literacy through platform design. We develop and design an intervention consisting of a **tweet trajectory**—to illustrate how information reached a user—and **contextual cues**—to make credibility judgments about accounts that *amplify, manufacture, produce or situate in the vicinity of* problematic content (AMPS). Using a research through design approach, we demonstrate how the proposed intervention can help discern credible actors, challenge blind faith amongst online friends, evaluate the cost of associating with online actors, and expose hidden agendas. Such facilitation of credibility assessment can encourage more responsible sharing of content. Through our findings, we argue for using trajectory-based designs to support informed information sharing, advocate for feature updates that nudge users with reflective cues, and promote platform-driven media literacy.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Collaborative and social computing systems and tools**.

Additional Key Words and Phrases: misinformation, trajectory, propagation, credibility, social signals, media literacy, critical thinking.

ACM Reference Format:
Himanshu Zade, Megan Woodruff, Erika Johnson, Mariah Stanley, Zhennan Zhou, Minh Tu Huynh, Alissa Elizabeth Acheson, Gary Hsieh, and Kate Starbird. 2023. Tweet Trajectory and AMPS-based Contextual Cues can Help Users Identify Misinformation. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 103 (April 2023), 26 pages. https://doi.org/10.1145/3579536

1 INTRODUCTION

The advent of new media technologies have made it possible for online accounts to engage in spreading and even manufacturing information of questionable credibility [47, 60, 89]. Often, such a spread occurs as users are unable to effectively assess the credibility of the source or they lack the complete context and reason behind the spread of the information. One such example includes the amplification of false information that paid protestors were being bused to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
https://doi.org/10.1145/3579536

demonstrate opposition towards one of the Presidential candidates in 2016; social media users were unable to assess the credibility of the source account with a mere 40 followers and contributed to sharing the posted content about 16k times on Twitter and 350k times on Facebook [58]. The Covington Catholic High School controversy is another example in which a video with misleading framing about a faceoff between a Native American man and a group of high school boys got about 2.5 Million views and more than 14k retweets [92]; the video, originally posted on Instagram with incomplete context, was popularized on Twitter in what was later identified as a deliberate attempt to spread the video by an account with misleading profile information that was subsequently suspended by Twitter [27].

Though efforts to detect amplification by bot accounts on these platforms may be useful [85], they do not address the perhaps more prominent role—as we see in these cases—of real human Twitter accounts in disseminating misinformation. How can we (as researchers and designers) empower users of online media platforms to thoughtfully engage and discern such problematic content from the larger pool of information—especially when sharing it. For example, the current retweeting scenario on Twitter 1 does not provide much insight (except the root, number of likes and number of replies) about the spread of the tweet.

Fig. 1. Screenshot from Twitter about the retweeting scenario when a user considers sharing a tweet using the 'Retweet' or 'Quote Tweet'. The platform provides little additional information on-hover about the context about the tweet.

To address the emerging need of curtailing the spread of problematic content, social media platforms—that only displayed popularity metrics in their early days—started introducing warning labels [54, 64, 93]. Additionally, there has been a push to use automatically generated social signals to indicate if a Twitter account might have shared misinformation in the past [44]. Following recommendations from others [18, 45], here we explore the potential of supporting media literacy with platform design, i.e., giving users more contextual information about the content they see, including where it originated and who brought it to their attention, to help them make better decisions about whether to consume, engage with, or reshare that content. We advance the scholarship on social signals and propose that providing **Contextual Cues** based on an account's recent online activity can provide insight into whether that account operates in good faith or not, and signal any potentially problematic behavior. Our chosen contextual cues provided participants with easy access to four types of information about a Twitter account: personal account-related information (e.g. account name, number of followers, etc.), account's activity in the past four weeks (e.g., hashtags used, accounts retweeted, etc.), distribution of account's activity (e.g, number of tweets, number of retweets, etc.) and age-distribution of other Twitter accounts that retweeted the account.

While the metrics to help users assess the merit of the content have evolved over some time, they have primarily focused on either the content or its initial source. Given the ease with which online accounts can share and make problematic content viral, we believe that these platforms should focus not only on the root (i.e. who authored the post/tweet) but also upon the actors involved in diffusing the content. Inspired by the concept of a trail [30], we propose

that exposing previously obscure accounts that popularized a tweet within the retweeting scenario—in addition to the known origin—along a **Tweet Trajectory** can help Twitter users to assess the credibility of the content in that tweet. Realizing the importance of different actors involved in spreading content, the proposed design of tweet trajectory consisted of the root account (that created the tweet) [66], a popularizer account (that popularized the tweet) [45], and a friend account (that may have liked the tweet, leading it into the feed of the user) [35, 84].

For guiding the choice of our contextual cues, we developed a framework to contextualize the different problematic behaviors exhibited by a Twitter account, i.e., *amplify, manufacture, produce and situate in the vicinity of* problematic content (AMPS). We designed an intervention consisting of a tweet trajectory (representing how the tweet came into the user's feed) and AMPS-based contextual cues (indicating past behavioral information) about the accounts involved in the trajectory. We adopted a research through design approach [34, 97] for this research and interviewed 21 participants—who were diverse in their political alignment and life experiences—using the intervention that we personalized to their individual tweeting experience and environment. We then analyzed the data using an affinity-diagramming-based thematic analysis. The contributions of this work are:

(1) We propose a design intervention modeled after the novel concept of a tweet trajectory and AMPS-driven contextual cues to reimagine the retweeting scenario.
   - By examining the concept of **AMPS-based contextual cues**: We demonstrate that exposing the tension between an account's suggested purpose and recent activity is useful for users to judge the authenticity of that account. Such a judgment can be useful to assess the quality of information shared by online actors in the context of their online associations [31, 52] and/or any other contentious behaviors. For example, inferring health-related information shared by a journalist as credible and meant *to inform* vs inferring radical content circulated by a partisan account as questionable and meant *to misguide*.
   - By examining the concept of a **tweet trajectory**: We demonstrate that designs that offer transparency about propagation of information can make users question their trust towards (inexpert) popularizers and (unfamiliar) friends that shared it, and make them reflective about the potential consequences of sharing it further. We use the findings to argue why resurfacing the role of institutional credibility obscured by the networked nature of social media [40, 63] can be useful to curtail the spread of misinformation.

(2) We suggest design recommendations for features that can provide users with reflective cues and help them become critical about their online feed and discern problematic content—including cases of mis/disinformation.

(3) The proposed intervention will guide future designs of intervention aimed at discerning misinformation and enable researchers to ask many interesting questions about facilitating credibility assessment and promoting media literacy within retweeting-like scenarios that need quick decision-making.

2 RELATED WORK

2.1 The need for new media literacy initiatives

Online platforms like Facebook, Twitter, etc. have made it easier to spread misinformation or to engage in inauthentic online behavior [4, 47, 60, 79, 89]. Some of the notable examples include how Internet Research Agency—a Russian entity that is known to orchestrate influence operations online—accounts gained their following and manufactured disinformation on Twitter [23, 56, 94], how internet groups manipulated and propagated selective news frames that initially surfaced on *8chan* [62], etc. In order to carefully navigate the potentially polarized and sometimes manufactured

information online, it is important for users to armor themselves with knowledge and skills to identify such inauthentic organized efforts [36] that are not uncommon to new media platforms.

While unpacking the meaning of new media literacy, Chen et al. (2011) argued that the move from traditional media literacy to new media literacy brings attention to users' ability to not only understand media content at the functional level but also to evaluate, analyze and question their understanding of it critically [15]. For supporting users in being critical of what information they consume in their environment, Ukraine in 2015 administered the 'Learn to Discern' information literacy program as a response to the flood of misinformation produced by Russia in 2014 and found it effective to help users discern the reliability of sources when presented with conflicting information [39]. Such media literacy efforts unfortunately have often been understood and promoted to be an individual's responsibility. As a matter of fact, the curriculum of the National Association for Media Literacy Education (USA) discusses five actions (Access, Analyze, Evaluate, Create, and Act) to be performed at an individual's level to become a critical thinker [7, 50]. Brodsky et al. have also found short term improvements [10] in the adoption of fact-checking strategies online by introducing lateral reading techniques that involve reading quickly but from multiple sources [91].

To develop media literacy efforts that are useful to identify modern-day information machinations, it is important to realize the limitations of individual disconnected experiences and to define and implement them at the level of an organization, platform, and nation [11]. In this research, we demonstrate how platforms can support media literacy initiatives by incorporating some of the features that provide the background context in which information operates and help users think critically.

2.2 New media signals to assess information quality

Social media platforms in the earlier days focused on popularity metrics like the number of likes and shares to guide users about engaging with the online content [42, 43, 57]. To address the emerging need of curtailing the spread of problematic content, the design of these platforms evolved to include warning and corrective labels that provide additional context as considered appropriate by platforms [53, 54, 64, 88, 93]. Preemptive labels aimed at inoculation have been effective to make people aware how they can be misinformed thereby increasing their resilience to it [55, 73, 87]. Corrective labels while usually useful, sometimes are known to cause a backfire effect in the users who come across the labels, i.e., strengthen their belief towards the misconception that the label is trying to rectify [5, 70, 82].

To support everyday platform users contribute and guide the application of these labels, in 2021 Twitter introduced a community-based approach called Birdwatch [19, 71] so that users can add context to a tweet. Such labeling unfortunately has been found to support partisanship rather than promote fact-checking [2]. Automated algorithms trained on human-labeled data about credibility of tweets have at times rendered users unhappy as they disagreed with the credibility labels due to individual differences of what and whom they considered as credible [38].

The continued push to include automatically derived credibility signals given the advantage that it is difficult for users to fake them unlike their profile information has been successful to a good extent [24, 38, 49]. For example, using labels that reflect accuracy of headlines have been found effective towards reducing the sharing of false information [48]. Researchers have demonstrated that automatically derived nudges can be effective towards providing some context into the credibility of information [6, 44]. The platforms-driven labels unfortunately haven't always been effective given that users have higher intent to consider verification as proposed by these nudges when they see that the message is congruent to their own ideologies [26]. In addition, these approaches are more suitable to the automatic way of thinking (over reflective) [12, 41] to support quick decision making. To help identify information machinations of a more

organized nature and preserve user agency, we believe that nudges that support reflective cues like the intervention that we propose can be more effective.

2.3 Information provenance, path and propagation

The source of information is considered to be one of the prime factors for assessing credibility of the content [46, 66]. Despite the challenges of identifying the information provenance and verifying its credibility for researchers and media platforms [20, 29, 79], cues about the source of information may not always be effective towards assessing content-credibility and detecting misinformation [21].

To benefit from other signals apart from information provenance, researchers have utilized diffusion-based metrics to help users assess credibility of the content [59, 72]. For example, Finn et al. developed a tool 'Twitter trails' to help users investigate a tweet of their interest if were be a potential rumor based on how it propagates within a social network [30, 65]. Shao introduced Hoaxy platform along similar lines to facilitate collection, detection and analysis of all incoming tweets to detect misinformation online [75]. While these platforms are extremely useful for assessing credibility, at present they are more suitable from an investigative perspective like that of a journalistic than for an every day social media user. Borrowing upon this concept, we introduce Twitter trajectories that illustrate the spread of information in a simpler fashion by highlighting specific actors involved in facilitating the spread of a specific piece of content from its source to the user.

Researchers are aware that information spreads online through a network [4, 78]. Network properties and behavioral features of propagation are useful for discerning misinformation from good information [67, 95]; e.g., witnessing simple characteristics of a Twitter account's social network (e.g., information about followers and following) can assist users in making informed decision about sharing content from that account [90]. Our approach exposes the behavioral traits and offers a much richer insight into the social network of multiple actors involved in putting information out there.

3 DESIGNING THE INTERVENTION

We adopted a research through design (RtD) approach for this research [34, 96]. RtD is a research method that facilitates discovery of new knowledge using methods of design practice [96]. Researchers have demonstrated the use of such an investigative approach to discover how and why variations in a design can impact users' affective and decision-making responses in different contexts [14], including that of misinformation [13, 77]. Along similar lines, we wanted to curate a set of cues that provide rich context about the spread of a tweet and encourage participants to explore *how and why* they could use such an intervention towards the credibility assessment of content when retweeting.

To help users understand the role of and context about different actors who are involved in spreading the information, we first came up with a framework that later guided our selection and design of the contextual cues. Next we chose the different actors to be shown in the trajectory that are important to explain how information reached a user while balancing the amount of information to be shown to a user. Our intervention consists of both the contextual cues and the tweet trajectory which we then use as a probe in the interview sessions. Given that a design intervention conveys a specific framing of the problem that we wish to explore, we note that knowledge generated through RtD is reflective of the functions and limitations of our intervention (as understood by the participants) [97] and as imposed by the researchers [25]. We now describe the process of designing the intervention—that we believe is novel, highly relevant to providing informational context, and extensible for other researchers to build upon [97].

Table 1. The proposed **AMPS** framework captures these four types of problematic behavior in which a Twitter account can knowingly or unknowingly participate.

Behavior	Description
Amplify problematic content	An account can bring more attention to problematic content by sharing it on their Twitter feed. For example, retweeting radicalizing tweets, content from polarizing accounts, or links from partisan media, etc.
Manufacture online dissent	An account can participate willingly in an inauthentic organized activity. For example, help to get an ambiguous tweet popularized out of context [92], partake in sponsored activities [33], etc.
Produce problematic content	An account can itself generate content that is harmful to the health of the online information space. For example, tweeting hateful speech or disinformation, using recent hashtags implicated in problematic activity, etc.
Situate itself within a problematic network	An account can be in close proximity to accounts that participate in any of the above three behaviors. For example, develop close network ties with problematic Twitter accounts by following them directly or getting access to their content through a shared connection, etc.

3.1 AMPS framework and contextual cues

For identifying the best cues that signal problematic behavior of a Twitter account, four of the authors who were students enrolled in a technology-design University program participated in a brainstorming exercise. Borrowing from existing research and our own experiences, we conceptualized multiple possible cues—without subject to the feasibility of operationalizing it—that signal if a Twitter account could be considered problematic. Examples of such cues from existing research include retweeting excessively (over other activities like tweeting) as it can signal amplifying behavior or 'pandering for social capital' [9]; using hashtags that consist of hate can signal producing radicalized content [1].

During the generative brainstorming process, we came up with 46 different (but not necessarily exclusive) cues that we believed can be used as a proxy for problematic behavior. Examples ranged from cues that are based on the tweet-content (using incendiary language, using recently adopted hashtags, tagging multiple popular people in their comments to grab attention etc.) to cues that are based on account-connections (following and/or retweeting from newly created accounts, exclusively following only highly popular/political accounts etc.). The exercise also surfaced the need for cues that should capture highly specific behaviors like creating a dedicated Twitter account to mirror the behavior of the parent account that it intends to mirror, only being active during polarizing events etc.

To make the most out of the RtD exercise and discover a wide range of insights, we intended to select a subset of these 46 cues such that they represent several problematic behaviors (as opposed to only one). Accordingly, we clustered the different cues based on the problematic behavior about which they offered insight. We then chose some cues from each of the four clusters described in Table 1. The AMPS framework characterizes four kinds of problematic behaviors—amplifying problematic content, manufacturing organized inauthentic behavior, producing problematic content, situating itself within a problematic network.

With a focus on operationalizing the different cues (corresponding to the four problematic behaviors discussed above) based on the information publicly available on Twitter, researchers then collectively discussed the feasibility of the cues that emerged during the brainstorming exercise. For example, displaying the two most frequently used hashtags used by an account seemed more feasible than displaying the two most hateful hashtags used by that account given the

Table 2. The different contextual cues that we included in the cue cards based on the recent 4 weeks of account activity using Twitter API v2. We also indicate the corresponding problematic behaviors from Table 1 that these cues intend to capture (A: amplify, M: manufacture, P: produce, S: situate).

Cues in the intervention based on 4 weeks of activity of an account	Related **behaviors**
Two most frequently used hashtags by that account	Produce
One hashtag used by that account along with another account that popularized it	Produce + Situate
Two most retweeted accounts along with the frequency of retweeting them	Amplify + Situate
Two most frequently shared domain names by that account	Situate
Frequency distribution of account's activity: tweets, quote tweets, retweets, & comments	Amplify + Manufacture
Distribution of other accounts by their Twitter age who retweeted this account	Manufacture

variability in their subjective interpretation. We also discussed opportunities to add more context with each cue to help participants better contextualize the information. For example, including which Twitter account brought attention to a certain hashtag provided context into how that hashtag might be used by others on the platform. Similarly, including the frequency of how often another account was retweeted by a Twitter account conveys the strength of shared beliefs between those two accounts [8].

Table 2 describes the different cues that we decided to include in the intervention. For each identified cue, Table 2 also indicates the corresponding problematic behavior enlisted in Table 1. It is possible that each cue can provide a signal into multiple AMPS-based behaviors, e.g., hashtag can indicate what kind of content an account produces and what community it situates within. Figure 2 provides more insight into these cues that we included in the intervention.

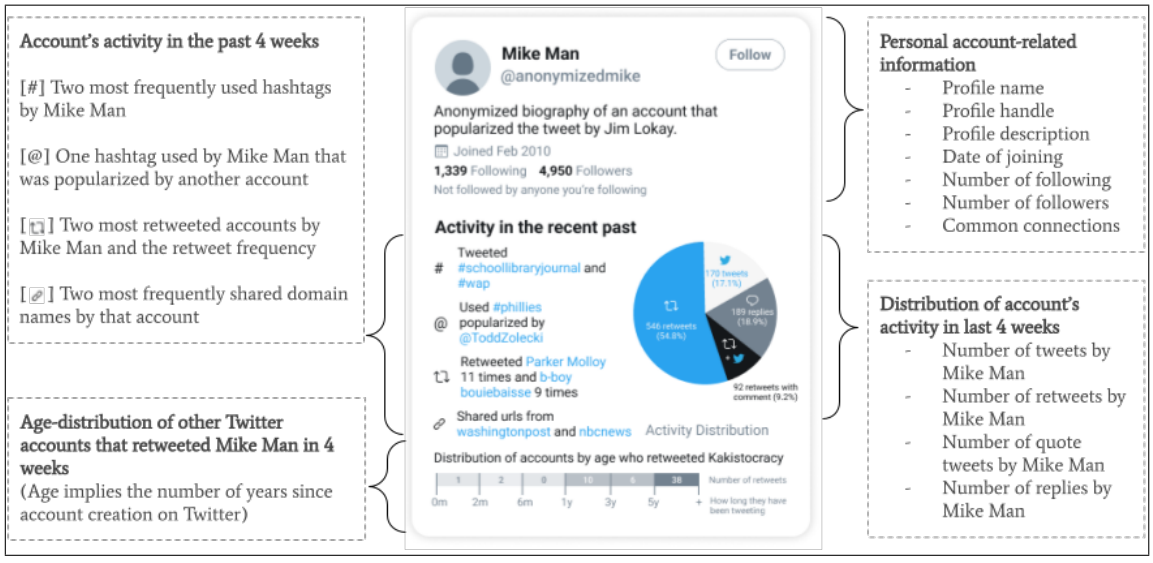

Fig. 2. Representation of a cue card (center) that we showed to a participant. Each cue card is populated with contextual cues specific to the activity of the specific Twitter account in the 4 weeks just before the interview session. The details (left and right) about these cues are mentioned alongside the card. Note that each cue might provide a signal into one or more of the AMPS-based behaviors.

Once we identified the cues of our interest that could be fetched out of publicly available Twitter information, we followed an iterative design process to come up with a suitable prototype that balances the role, look, and difficulty of implementation of the prototype [97]. The final proposed design also benefited from feedback that we gathered through a user-research activity from five graduate students enrolled in a University-based technology-design program.

Aligned with the recommendations of Aspen Institute's report 'Commission on information disorder'—particularly about empowering users through digital interventions that give them the skills and context to safely navigate low quality [45]—we offer our participant automated cues out of readily available Twitter information that can provide the necessary context about the information with which users might be engaging. We believe the proposed cues bring attention to the much needed human aspects of information propagation [28], i.e., the personal motivations of why different accounts may post or share certain information.

3.2 Tweet trajectory

Finn et al. introduced the concept of Twitter trails that offers several visualizations to help users investigate how a particular story originated and then propagated within a social network through a series of tweets [30]. Following up on the concept of investigating information propagation and adapting it to quick-decision making that occurs in the retweeting scenario, we introduce a Tweet Trajectory 3 to demonstrate how a tweet could have reached the user's Twitter feed. We particularly focus on Twitter accounts that were responsible for bringing widespread attention to the tweet (i.e., popularizers of that tweet) and accounts that serve as a connection between these popularizers and the user considering to retweet that information.

Fig. 3. Our proposed "tweet-trajectory" consists of the root tweeter account, a popularizer account, and a friend account.

We chose our first actor to be the known source of the tweet (root tweeter) given the established significance of source for online credibility assessment [35, 66]. Realizing the potential of top accounts with large number of followers in spreading content [17, 45], we chose our second actor to be a popularizer of that information. Given that users tend to trust online information more if it is shared by one of their friends [35, 84], we chose an individual-user's online friend as our third actor along the trajectory. By situating the proposed trajectory-based intervention within the retweeting scenario itself, our approach benefits from the realization that users tend to skip investigating the credibility of online content given the need for extra effort [35]. By displaying the AMPS-based contextual cues in a trajectory, our intervention (Figure 4) aims to help users identify and understand specific manipulation tactics—which is a critical requirement of new media literacy [74].

4 STUDY

4.1 Participants

To encourage the discovery of a wide range of techniques that users employ when assessing the credibility of online content as facilitated by the AMPS-based cues and trajectory, we recruited participants that were diverse in terms of their political alignment and life experiences. To ensure such diversity, our recruitment survey asked interested participants about their trusted media channels that serve as their everyday source of information, level of education, and (urban or rural) neighborhood. The survey also asked participants about their Twitter handle to confirm if they had an active Twitter account for the last six months or more. Some prior experience with the Twitter platform was essential so that the participants can comprehend and reflect upon the information in the AMPS-based intervention.

We began posting the call for recruitment in multiple online spaces like Twitter (publicly accessible), Facebook (closed groups about specific media personalities), and SurveySwap. Given the asymmetrical political alignment of participants who expressed interest through these channels, we next posted the same recruitment survey on the MTurk platform [69]. Upon realizing that most of the interest from the Mturk platform included participants without active Twitter accounts, we next posted the recruitment survey in the Craigslist-volunteers' section in several American cities that we adjusted according to the need for diversity of our participants.

We recruited 24 participants (out of 90 expressions of interest) that meet our inclusion criteria and helped to maximize variance in political alignment and life experiences. We sent the consent form with details about the research to these participants. Of these 24, three participants opted out, citing discomfort about sharing their thoughts on how and why they choose to retweet something on their Twitter feeds. Out of the 21 interviewee participants, 10 were living in an urban, 8 were living in a suburban, and 3 were living in a rural neighborhood. The median following of our participants was 472 (min: 99, max: 2990) and that of their followers was 369 (min: 16, max: 29990). Our participants, who ranged from 18 to 60+ years of age, referred to a diverse set of media sources for their trusted news including Washington Post, New York Times, Fox News, The Hill, MSNBC, and others. Table 3 describes participant characteristics in detail.

4.2 Personalizing the intervention for each participant

To encourage participants to discuss how they might employ the AMPS-based intervention in a retweeting scenario, we personalized the intervention to closely approximate their experience on Twitter.

Selecting the tweet: For every selected participant, we identified a possibly-contentious and relatively popular tweet which the participant had retweeted in the recent past (relative to the participant's Twitter activity). To confirm that the participant did not see this tweet directly without the agency of other accounts, we ensured that the participant did not follow the Root tweeter.

Selecting popularizer(s): Using the Twitter API v2, we fetched the most recent 100 retweeters of the tweet. Out of these 100, we then selected 3 retweeters with the highest number of followers. We used these three popularizers in 3 unique tweet trajectories that we showed to the participants.

Selecting the friend: Once we selected the popularizers, we then chose a Twitter friend account—i.e., an account which follows the participant, and the participant follows it back—randomly out of the recent 5 accounts with whom the participant had any Twitter interaction (like, retweet, comment, quote tweet). We also confirmed that the friend account does not directly follow the Root tweeter, implying that it was only through some Twitter account that the friend came across the tweet.

Table 3. Characteristics of the interview participants as captured through a survey.

P#	Following/ Followers (count)	Age (yrs)	Neighbor-hood	Education	Tweets/ Retweets (weekly)	Media sources (primary)	Regretted sharing a tweet
P1	1301/351	35-45	Urban	Bachelor's	5+	Buzzfeed, CNN, Haertz, TechCrunch, Seattle Times	1+ times
P2	461/611	35-45	Rural	Bachelor's	2-3	BBC	Never
P3	1098/2474	35-45	Urban	Graduate	5+	WaPo, Washington City Paper, DC Line	1+ times
P4	1613/1265	35-45	Rural	Graduate	5+	NPR, Scientists, PBS, CNN, Health Experts, Journalists	1+ times
P5	201/121	25-35	Suburban	High school	5+	Seattle Times, Trending topics	1+ times
P6	1360/487	35-45	Suburban	Graduate	2-3	NYT, The Hill, NPR, Journalists across outlets	1+ times
P7	434/366	25-35	Urban	Bachelor's	5+	WaPo, NYT, WSJ, Fort Worth Star-Telegram	1+ times
P8	478/369	18-25	Urban	Graduate	5+	No specific media sources, ACLU account on Twitter	1+ times
P9	1305/782	35-45	Suburban	Bachelor's	once	Fox News	1+ times
P10	342/899	18-25	Suburban	Some school	once	NYT, WSJ	Never
P11	173/29900	25-35	Urban	Bachelor's	5+	NYT, The Atlantic, WSJ, Quillette	Never
P12	99/16	45-60	Suburban	Bachelor's	5+	Fox News, CNN	1+ times
P13	1334/438	35-45	Urban	Bachelor's	5+	Fox News, MSNBC	1+ times
P14	2990/2582	60+	Suburban	Bachelor's	5+	MSNBC, CNN	1+ times
P15	121/41	18-25	Urban	Some college	once	Individuals I follow, Twitter news/trending	1+ times
P16	2555/2791	45-60	Urban	Bachelor's	rarely	MSNBC	1+ times
P17	110/71	35-45	Suburban	High school	rarely	Fox News on Twitter	Never
P18	472/262	18-25	Suburban	High school	5+	CNN, Congress members, Twitter trending	1+ times
P19	450/91	25-35	Urban	Graduate	once	Telesur English	1+ times
P20	256/175	25-35	Urban	Graduate	5+	Reddit news/trending	1+ times
P21	986/173	45-60	Rural	Bachelor's	2-3	NPR, MSNBC, BBC	Never

Populating cues: For every Twitter account in the trajectory (root, popularizers, friend), we populated the respective cue card based on their publicly available Twitter profile information and their publicly accessible Twitter activity as indicated in 2.

Upon the failure of any of the above criteria—e.g., chosen friend account follows the root tweeter—we selected the next choice that fits our criteria for personalizing the intervention.

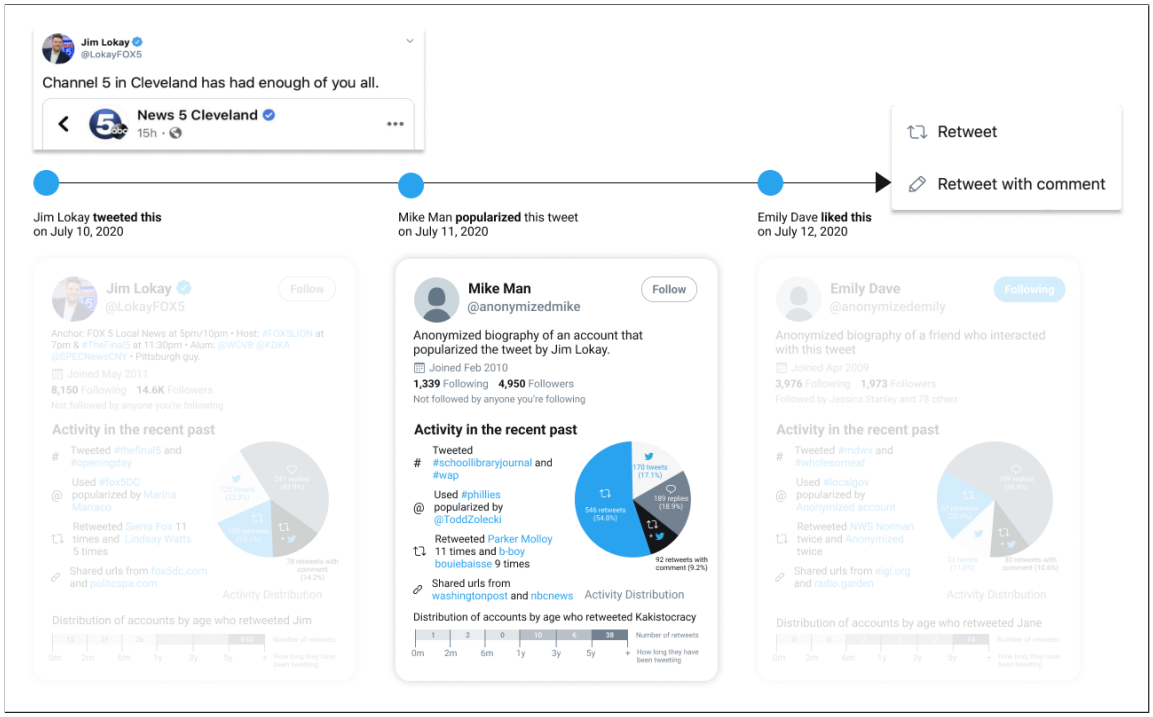

Fig. 4. Reimagined retweeting scenario using AMPS-based cues and tweet trajectory corresponding to Jim Lokay's tweet personalized for a participant of the study. Jim Lokay here represents the root, Mike Man represents the popularizer, and Emily Dave represents the friend who may have liked and brought this tweet into the participant's Twitter feed. All the accounts in this intervention are either verified by Twitter or anonymized to protect privacy. Every participant saw three of such trajectories with varying popularizers in between the same root and friend.

4.3 Interview procedure

Each interview lasted about 60 minutes and involved one participant, one researcher to guide the discussion, and one researcher to take notes. We conducted the session online over Zoom and compensated the participants with $20 each towards an hour of their time.

In the first part of the interview, a participant shared their Twitter screen and walked us through their Twitter feeds. During this walkthrough, we encouraged them to talk about why they might or might not share any tweet that was in their feeds to understand how they thought if any information was worth sharing. We also asked them what all cues do they refer to when making such a decision.

In the second part, the researcher shared their screen and introduced them to the concept of cue cards (as illustrated in Figure 2) in four steps. In each step, we showed them one of the four sections and confirmed that their understanding of the information in the cue cards aligned with the intent of its design. We also encourage participants to think and share with us how the information in each of these cards might be useful if they were thinking about retweeting from the account whose cue card was shown.

Next, we showed them the tweet that they had retweeted in the past and that we chose to populate the cues in the personalized AMPS-based intervention. We asked them to (re)imagine the scenario in which they are considering

retweeting the tweet and to describe in as much detail as possible how they decided to share that tweet when they first saw it. We note that some of the participants were not too certain if they had shared the tweet citing that retweeting is a snap decision.

Next, we showed them one of the trajectories of that tweet as illustrated in Figure 3 and asked them what the illustration means to them. We then showed them our designed AMPS-based trajectory (Figure 4) and asked them to talk aloud about how they might use (if at all) the information in the intervention to guide their decision-making. We offered them two more choices about the popularizers in the trajectory and discussed which of the popularizers they thought impacted the credibility of the tweet and their decision of retweeting it.

4.4 Data analysis

We conducted an interpretive, grounded analysis of the data that was collected through participant interviews in the RtD exercise. First, each researcher who conducted a participant interview transcribed that session. Another researcher and the lead researcher then went through that entire transcript separately to check for any corrections and to get familiar with the gathered data. Transcribed interviews were then atomized into cards which were then organized using an affinity-diagramming approach. Through the analysis, we identified and clustered the common themes as they emerged across our participants to discover how the intervention—comprising of the cues and the trajectory—supports users towards credibility assessment of a tweet. We iteratively refined these themes and recorded the insights using analytical memos. In section 5, we report on how contextual cues support participants in developing and/or refining their mental model about an account. In section 6, we then report insights about how the tweet trajectory complements these efforts towards making a quick-but-informed credibility judgment.

5 FINDINGS: HOW AMPS-BASED CONTEXTUAL CUES (CC) SUPPORT CREDIBILITY ASSESSMENT

When assessing the credibility of a tweet in the present design of Twitter, participants often mentioned that they go to a Twitter account's profile to check what and how they contribute to the online discourse. Upon being introduced to the cue cards designed after the AMPS framework, participants found the provided summary of an account's behavior in the past four weeks to be effective towards the rapid nature of decision-making in the retweeting scenario.

> "It adds to the heuristic of good information because you are giving me a snapshot of what they have done and how they did it, rather than me clicking on the guy's profile, seeing the top section of the card and then scrolling legitimately through this person's feed to see if they retweeted a lot, are they commenting a lot. This gives me a snapshot and shortens the amount of research that I have to do." - P2

We now report how participants employed the different cues in the cards towards building and/or refining their mental model of the different accounts and made a credibility judgment.

5.1 CC1: Authenticity of an account

5.1.1 Profile description helps to infer the purpose of the account. When assessing the credibility of a tweet, participants often mentioned looking at an account's profile information to get a signal about the user and to infer that account's purpose. While Twitter offers this information—profile picture, profile name, profile handle, profile description, number of followers, number of following, common connections—only on-hover, the cue cards that we designed surface and highlight this profile information within the retweeting scenario. Participants found the higher salience of profile

description useful towards identifying the account's purpose for being on Twitter—and anticipate what kind of content it might *produce*.

> "I know that <NAME> works—I mean she says right there—for Governor <NAME> as their contractor. And so from that, I would surmise that she's neck-deep in COVID-19 response stuff right now. I think I like her. I would imagine that her tweets are a combination of democratic political views." - P6

In the example above, the participant used the profile-related information of an unfamiliar account to infer that account's occupation. While in this case profile description added credibility to the tweet, in other instances, it discredited the tweet. For example, in the tweet below, the participant inferred that the account possibly wants to spread pro-life content and decided to not engage with it further.

> "I cannot get past the account saying that murder is their business and then their profile picture! Makes me nervous that people like it. It looks like pro-life propaganda." - P9

5.1.2 Profile description and account activity collectively facilitate a consistency signal to help infer account authenticity. Once participants set their expectations about the type of content to expect from an account, they used the activity distribution for deciding if that account's behavior is true to its purpose as inferred by them. This search for consistency between account purpose and account activity helped participants ascertain whether an account is authentic, i.e., is the account what it claims to be. Higher perceived authenticity of an account implies more credibility [61] and might suggest lower chance of any compromised activity like *manufacturing* dissent. When doing a consistency check, participants associated different behaviors with different account purposes. For example, they associated replying relatively more (than tweeting) with regular users' behavior on Twitter, but expected an influential account to generate more original tweets.

> "If it is a politician, I would like to see that they have a lot of their own tweets. Or let us say if it were someone with a large platform, I would want to see they are spreading their word." - P10

This search for consistency within a card to ascertain authentic behavior sometimes led to discovering unexpected activity, not only from unfamiliar accounts, but also from accounts that were in their following network. Given that participants had prior mental models of familiar accounts (unlike unfamiliar accounts), they switched their attention from *is an account what it claims to be* to *is an account what I thought it to be*. In a few instances, participants were surprised to see some of their friend's activity when going through their friend's cue card, e.g., in some cases, participants felt the activity of their friend—with whom they otherwise shared positive experiences—to be offensive. Such a discovery made them question the kind of faith they had in accounts with whom they shared strong ties on Twitter.

> "I don't know. It's like all of sudden the nature of what they are talking about and doing is not very appealing. I might decide to not follow them anymore." - P21

5.2 CC2: Online associations of an account

5.2.1 Hashtags used by an account can suggest shared values and interests. One of the signals that participants searched for when looking at a Twitter account was whether they had any common topic of interest with that account. The few hashtags displayed on cue cards were often enough to give participants some insight into the account's interests. Having a common interest with an account indicated that they shared a similar understanding of that topic—e.g., "If they are tweeting *Black Lives Matter*, then I'd be like cool, we are at least clear on that and not debating it." (P5)—or were involved in the same community—e.g., "She's a Patriots fan too" (P21). Discovering a common interest through the

content *produced* by an account that participants saw in the cue cards was largely seen as a sign of relatability towards the unfamiliar account.

> "I feel like she's very relatable because she's a mother, into traveling, and loves Disney. I am obsessed with Disney. She's like promoting Marvel movie's hashtags. We are very into Marvel in our house. So it's just like I feel connected because we have a lot in common." - P9

5.2.2 Retweets by an account and retweets of an account provide context into the broader community. Although participants used hashtags to get some sense of shared values and interests with another Twitter account as reported above, the extent to which this cue was useful to identify accounts that *produce* concerning content largely depended on the context in which the hashtag was used. For example, the inference about which community an account is active would entirely change based on whether that account used a political hashtag *#President* for expressing a sentiment of support or opposition towards the President.Participants found the information about Twitter accounts—who retweeted a certain hashtag and popularized its usage—useful to understand the context in which the hashtag is used and to identify the larger shared community around it, i.e., how the account is *situated*. In the example below, the participant used this information to judge if the accounts who retweeted that hashtag are from their community and if they approve of the context in which the hashtag was used.

> "If you can recognize the person who retweeted a hashtag, then you can get a better sense of where this information comes from or who is the voice behind the hashtag. You will then have to decide if you are a part of the community or outside of it depending on that person, but it will still become a breadcrumb for me to think about 'who is this Katie. I do not recognize her'." - P4

It is important to note that while most participants found the cue useful to identify accounts that *produce* content that they disapprove, such a signal could cause concern by encouraging selective and ideologically aligned engagements (described more in Section 7.4.2). In addition to indicating who popularized a certain hashtag, cue cards also displayed how often and which Twitter accounts were retweeted. Participants found this information—about two Twitter accounts that received the maximum number of retweets by the cue card account—useful to get a sense of the broader community of that account. For example, in the tweet below, the participant witnessed that the Twitter account in the cue card retweeted President Trump multiple times thus *amplifying* him. Participant then used this information to assess that the account, that was otherwise unfamiliar to them, might be *situated* in the pro-Trump community and decided to not associate with the account.

> "It could be helpful if I am looking into someone, and I don't know what their deal is, and then I see that they retweeted the President five times. That will be an example where I will be like *Oh God! Okay. Let us get out of here.*" - P19

5.2.3 URLs embedded in tweets are useful to identify one's information sources. The cue cards that we provided participants highlighted URLs embedded in tweets that the Twitter account in the cue card shared. Participants employed this direct access to the URLs for inferring where the information might be coming from to (infer how that account is *situated* and) question that account's overall judgment. Unfortunately, this was only possible when participants recognized the URL. In cases of ambiguity about making a URL-based judgment, a participant wanted to dig deeper into the URLs themselves:

> "If she has shared URLs from like Drudge Report or something, I might question her judgment a little bit. If she is doing WaPo and Drudge Report, it would raise a flag; I'd then want to see what exactly she shared from them." - P6

5.3 CC3: Contentious behavior of an account

5.3.1 Excessive activity of one kind suggests questionable behavior. When glancing at the activity distribution of an account as illustrated by the pie chart in the cue cards, most of the participants shared a similar sentiment that a high amount of retweeting-activity signals unreasonable and potentially toxic behavior of *amplifying* content on the Twitter platform. For example,

> "I think the distribution is the piece that matters more to me. This person is just a retweet machine without any care if the information might be coming from some kind of a nefarious product." - P4

A few participants also related such retweeting behavior to "shouting in the void" (P19) or to not putting much care when sharing information on Twitter. At the same time, they also acknowledged how it will be odd to see that an account's primary activity involves tweeting since a huge part of Twitter is to retweet others. As previously noted, a few participants found this cue about *amplification* only useful when considered together with an account's purpose to be on Twitter. For example, this cue did not signal any concerns in case of an activist-like account with a relatively high number of retweets as it could indicate the activist's passion to share useful information.

Participants also paid attention to the number of replies that the account has engaged with in the past four weeks. A majority of participants associated a higher number of replies or comments with the argumentative nature of that account. Upon witnessing the disproportionately higher number of replies than that of tweets, a participant expressed:

> "It would give me some pause to see that they never tweeted but replied a lot. That indicates that this person may just like to troll bigger accounts or maybe get into arguments or something like that." - P11

5.3.2 Too many retweets by recently created accounts suggest coordinated activity. Researchers have witnessed organized inauthentic activity [68, 76] on Twitter in which several questionable and often recently created accounts retweet a specific tweet from another account to amplify a tweet or an account that could be misleading or taken out of context. When inspecting the information about the distribution of accounts (by their age on Twitter)—who retweeted the account whose cue card was presented to them—participants were skeptical of an account that received too many retweets by newer accounts and found the information useful for inferring 'bot-like' or organized activity. Understanding how tweets that receive retweeting from bot-like accounts may be problematic is useful to suspect an attempt at *manufacturing* dissent.

> "I didn't realize before seeing this distribution chart that things like this can show if some tweets are misleading or that accounts might be trying to potentially mislead (by retweeting this account). Even if just as a warning, this will be useful to know if the account is engaging in good faith or are they just trying to pick up my time." - P5

This was the only cue in the cue card that indicated the kind of activity others exhibit related to this account. Though we presented the distribution of accounts in a highly granular fashion (e.g. separating three-year-old accounts from five-year-old accounts), most participants found this cue to be equally meaningful and easier to comprehend if it simply made a distinction between newer suspicious accounts and older established accounts.

6 FINDINGS: HOW TWEET TRAJECTORY (TT) COMPLEMENTS CREDIBILITY ASSESSMENT

The trajectory part of the intervention (Figure 4) consisted of three cue cards corresponding to the root of the tweet, the popularizer of the tweet, and the friend that may have liked and brought the tweet into the participant's feed. When reflecting on these three accounts in the tweet trajectory—that was personalized for the participant based on their Twitter network—users primarily asked three unique questions of them. They used the profile-related information and the activity distribution in the cue cards to see if the root is a content creator. When it came to the popularizer, users focused more on the popularizer's activity—i.e., use of hashtags, shared URLs in the tweet, etc.—to evaluate if the popularizer had the necessary expertise to weigh in on the original tweet; if not, users employed the same information to confirm the absence of any ill-agenda or the possibility of making a disreputable association. When it came to evaluating the friend cue card who brought the tweet in their feed, users mainly asked themselves if they could trust their online friend's judgment. We now report how participants used the different cues in trajectory to gain insights about how information propagated to them, and how information could propagate to others if they were to share it.

6.1 TT1: Trust towards information spreaders

6.1.1 Unfamiliar/surprising activity invokes skepticism in the participant's following network. Most participants indicated that their friend—i.e., the rightmost card in the trajectory—liking or retweeting content is similar to their "friend signing off on a piece of information". When comparing the trajectory-based retweeting interface with the existing retweeting interface, some participants however started questioning the trust that they had in their friends. For example, one participant attributed it to the relatively unexpected large number of replies made by their friend as inferred from the frequency distribution of account's activity in the cue card. Noticing either unfamiliar or unexpected activity in the trajectory made participants question their blind faith in their Twitter friend as expressed in the following quote:

> "If I trust the people I follow and they trust the people they follow, then we are in this strange circle of Twitter trust that is too good to be true. But having some of these unknowns in here made me stop a little. Having someone in the middle of the trajectory actually took something away based on how much I initially trusted the information because my friend was kind of vetted." - P2

The tendency to question accounts with whom participants shared strong ties—as also exhibited by the discovery of unexpected behavior inconsistent with their understanding of that friend account—was also affected by how the popularizer was connected with the friend account. When looking at the trajectory, participants thought about how the information reached from the source to the popularizer and, in particular, how it reached from the popularizer to their friend. They wanted to see the inter-personal relationship (e.g., do they follow each other) between the different accounts involved in the trajectory to better understand how the tweet traveled across different accounts:

> "I think I will be curious as to how this reached my friend's feed. I think what is like tripping me up about the middle popularizer card is that they seem like a random person. I wish I could tie their identity a little bit more concretely to someone that my friend follows or to someone who has a history of popularizing such tweets." - P8

6.1.2 Expertise of early popularizers imparts confidence in the information. When talking about how tweet-trajectories might alter the perceived tweet's credibility, participants often mentioned searching for expert voices in between the root tweeting the content and their friend liking it. Knowing that someone with expertise—whom they trust—shared the tweet early on in the trajectory positively impacted the participants' decision to retweet the content (as opposed

to noticing any ill-intent or Propaganda). The example below demonstrates how the self-disclosed information about being an academic in the description could be useful to make a judgment about the account owner's potential expertise.

> "I don't know these people but that the popularizer retweeted Matt Stolen rings a bell. The popularizer (based on profile information in the cue card) seems to be an academic, faculty at a renowned University, runs the lab for social media so presumably a social scientist. Then his shared hashtags, shared URLs from academic websites like cambridge.org also give me more confidence in the tweet." - P15

In addition to the profile description, the participant also used the additional information in the cue cards to support their judgment about the account's expertise. For accounts where it was not possible to infer their expertise based on self-disclosure, participants searched for any topical match across the account's activity in the cue card and the topic of the tweet. They associated a sense of expertise with accounts that shared information about a topic in which they had some prior experience and trusted them relatively more. For example, when it comes to information about medicine, a participant expressed how prior experience with the topic and history of sharing content about it might serve as a signal of credibility.

> "I see they are an MD. It makes me respect them when they tweet about medicine and public health. If it was someone else, I would be looking for their credentials in that space. If not, MD is not essential for sharing information on medicine, but have they written extensively about it? E.g. good journalists writing many evidence-based stories about it. If it was anyone saying 'drink bleach, it will cure you' is not the same as a Doctor saying it." - P6

6.2 TT2: Consequences of sharing information

Trajectories made participants wary of how their retweeting could help the information propagate further. By witnessing potentially disreputable accounts in between the root tweeter and their friend, participants sometimes grew hesitant to share the information—not because of the information itself but—because they did not want to get involved with the accounts who shared that information. Exposing potentially disreputable popularizers (inferred using hashtags they used and domains/accounts they shared) assisted users to avoid associating with them mostly from the perspective of preserving one's online impression. In the example below, one of our participants suggested *The Lincoln Project* to be one such account on Twitter.

> "I would probably be less likely to retweet if it was an opinion from an outlier of a source and I agree with it. I might like it, but I'm less likely to retweet it. And depending on this (popularizer-related) information, I might not even like it. The Lincoln Project is a really good example. I love what they are doing and I fully support their Republican-led revolt against the President. But I am never going to like or engage in it because at the same time I want to support the organizations that have been calling it out. So the Lincoln project is great for my dad, but it's not something I'm going to get involved in." - P6

The Lincoln Project foundation was founded in late 2019 by some of the incumbent Republicans as a part of their efforts to oppose the reelection of Donald Trump (the acting President then). However, their potential links to the pre-Trump GOP and adoption of questionable strategies for promoting the project was controversial among both the political camps. Given its controversial nature, the participant—despite supporting the larger message—wished to maintain distance from the account for impression management. A few participants with relatively high-media literacy in the study were wary that such an intervention would expose their role in propagating the information to

others and hence wanted to avoid such an association through retweeting. Researchers have witnessed this tendency of engaging with information selectively to manage one's reputation in the context of correcting their online behavior [3]. This example suggests how trajectories can be particularly useful in ambiguous instances of whether to engage with information that one sees online, or not.

Participants also attributed some of their past decisions about deleting their retweet (of another tweet) partly to the information itself and partly to their unwillingness to be associated with the account that shared the tweet. When discussing their past negative experiences about retweeting on Twitter, participants often mentioned the need to only associate with accounts that represented them well. A participant mentioned an incident where they deleted a tweet because they no longer wished to endorse the account that had originally posted it.

> "There was a time when someone posted a tweet that was critical of covid precautions when football season was starting given that the players get tested every day. I don't know if they were being sarcastic and I could not tell. But I retweeted it thinking that it was criticizing it from the perspective that it was dangerous and irresponsible to start football. Later I realized it was someone who was anti-science. I immediately deleted it as I didn't want to endorse the person." - P19

7 DISCUSSION

Information that we come across online in everyday life can be highly contextual [27]. For discerning problematic content from good information, it is important to identify and understand the context in which information is posted online. For understanding the context around information, we need to be informed about its provenance (*i.e.*, who started it) and its spread (*i.e.*, who shared it). While identifying the provenance of information can be tricky, social media platforms could do more to help users realize how different intentions of those who share information might shape the larger context around it and compromise its credibility. We address this missed opportunity through this research and introduce the 'AMPS' based intervention consisting of contextual cues and tweet trajectory. The intervention provided users information that—though publicly available on Twitter—users had to seek out and synthesize which is not feasible within the current design of retweeting feature. We now discuss the findings (summarized in Table 4) how the proposed intervention is effective for helping users reflect on their friends and their tweeting habits, thus making them more careful about what they share online.

7.1 AMPS-based cues help assess overall credibility of an account

The first part of the intervention involved contextual cues that signal four inappropriate behaviors about an account: amplify, manufacture, produce, and situate in the vicinity of problematic accounts. Using research through design as our method, we discovered that our contextual cues helped support users in assessing credibility in three ways. First, our design helped them assess account authenticity. Although amplifying by means of excessive retweeting was largely looked down upon, the extent of it being problematic depended on the inferred purpose of that account. Knowing the purpose of an account helped participants better anticipate the kind of content that account might *produce* or choose to *amplify*. For example, based on account description and activity, P7 inferred the account to belong to someone who "works in the news' and decided to trust the information they shared. A mismatch between the potential purpose as inferred by users and the content posted by that account could signal inauthenticity—making users more critical about the account and its content.

Our contextual cues also helped users assess the online associations of an account. Participants perceived hashtags to be useful for witnessing any problematic content an account *produces* provided the context in which these hashtags were used was clear. By considering such created content together with the accounts whom they retweeted and URLs which they shared, users were able to assess online associations of that account and realize what kind of community that account is *situated* in and/or the kind of content it might *amplify*. Users could then employ these judgments to evaluate the information shared by the actors in the context of their information neighborhood [31, 52]. For example, P2 discredited information from an account since they suspected the account "to have some propaganda" based on the activity-related cues and whom they have actively retweeted in the recent past.

In addition, our contextual cues helped identify contentious behavior. The relative distribution of an account's activity, i.e., how much it tweets vs replies etc. as seen in the cue cards was useful to infer any toxic behaviors of an account. Participants were also quick to realize that retweets from recently created Twitter accounts could also contribute to that account's contentious behavior. Such a concern combined with any previous suspicion based on account's inauthenticity hinted at possible participation in organized behavior to *manufacture* dissent. The AMPS-based contextual cues thus enabled users to assess the credibility of an account by employing context—not readily available on Twitter—that was offered by the cue cards.

Table 4. **AMPS-based contextual cues** help users to assess information credibility based on account's authenticity, online associations, and contentious behavior. **Trajectories** help users to assess if they should trust the propagators of information and should it be shared further.

Finding	Cause of concern	Example scenarios
CC1: Authenticity of an account (5.1)	Questionable **purpose**	Self-described or user-inferred propaganda account
	Inconsistent **participation**	Journalist account with no original tweets
CC2: Online associations of an account (5.2)	Lack of **shared interests**	Hashtags used in a disagreeable context
	Unfavorable **connections**	Retweeting highly polarizing accounts
	Uncredible **media sources**	Sharing media from toxic sources
CC3: Contentious behavior of an account (5.3)	High **#retweets/#replies**	Excessive replying suggesting argumentative behavior
	Getting **retweeted by new accounts**	Getting most of retweets from accounts that are one month old or less
TT1: Trust towards info-spreaders (6.1)	Unfamiliarity in **one's network**	Surprising political stance of a high-school friend
	Lack of **expert popularizers**	Movies-related account spreading vaccine content
TT2: Consequences of sharing (6.2)	Potentially create **problematic association**	Retweeting creates new cues (online activity) that are not representative of account's history and/or purpose

7.2 Trajectory helps assess credibility of information as it propagates through a network

Users employed the trajectory part of the intervention to aid their judgments based on the cue cards as reported in Section 6. First, the trajectory made users evaluate their trust—an important factor for credibility assessment [51, 80]—not only towards an account but towards the larger network of information spreaders. Upon discovering surprising elements about their friend—e.g., discovering that their friend has changed their political interests over time (P12)—users

grew skeptical of why their friend shared the content from the popularizer and if they did any due diligence prior to sharing. This sudden sense of unfamiliarity towards their friend made users question their knowledge-based trust in their own network. Knowledge-based trust deals with the ability of the person to predict a person's behavior based on their past experiences. The easy access to background actors who spread information afforded users an opportunity to notice the absence, if any, of expert popularizers and also that of any social or organizational structures that can bring institutional credibility when assessing the quality of information. Thus, the trajectory impacted knowledge-based and institutional trust [16, 37] in a way that is not facilitated by the present design of Twitter.

At times, trajectories surfaced tension as users found information that they supported to be circulated by actors with whom they did not want to associate. When considering the consequence of sharing this information, users evaluated the cost of such a problematic association and sometimes decided against it. Thus, trajectories also impacted calculative trust that is based on rewards and penalties [37].

The present design of social media—as Hancock describes [40]—is based on providing content produced by one's online network thereby increasing trust amongst different connected actors, i.e., one's following on Twitter; this in turn reduces the institutional trust in organizations like government, academia, scientific groups, etc [63]. In other words, users tend to trust low quality content shared by their trusted friend more than high quality content shared by distant friend [81]. We believe that by resurfacing the role of institutional trust obscured by the current design of social media, and having users re-evaluate their knowledge-based trust in their network, the proposed intervention can be useful to question one's trust towards information spreaders and thus curtail the spread of misinformation.

7.3 Using AMPS-based cues and trajectory beyond Twitter

Although we situated this study in the context of Twitter platform, the findings and implications of this research are relevant to several online media platforms. The merit of any information on such platforms is not merely a property of the content itself, but that of a network of online actors in which the information spreads. By making all the actors who participate in sharing content (using trajectory) and the context in which they engage with that content (using AMPS-cues) salient, the proposed intervention or its variations can enable users to assess information not merely based on the content but also based on the larger ecosystem that promotes it. In crucial times when these ecosystems rapidly evolve through organized and coordinated efforts with an ill-intent, such transparency into information spread can help users think through the implications of sharing content in ways that are not permitted by current platform designs.

Information operations that can be seen on online platforms are known to be highly networked [56]. However, techniques to deal with misinformation—e.g., content moderation—do not benefit equally from such networked-ness [22]. In addition, false stories are known to spread about six times faster than true stories [86]. By communicating networked-ness of information to more users, trajectory-based interventions could facilitate effective decision-making and curtail the rate at which misinformation spreads.

This study demonstrates our interventions can effectively communicate how information spreads online and thereby help communicate information credibility as property a larger network of evolving accounts. By communicating the machinations of information propagation to everyday users, such a mechanism can help platforms to promote information literacy that is much needed for engaging with online content responsibly.

7.4 Design implications to encourage responsible information sharing

7.4.1 Provide cues that signal account's evolving behavior. The findings of the study made it clear that social media platforms need cues that reflect the real nature of an account's everyday activity so the account cannot easily lie (e.g.,

in their profile description) without inviting scrutiny from other users. Our proposed intervention could help users identify several causes of concern when it came to sharing information enlisted in Table 4 and made them question if the information shared by that account is trustworthy. While the intervention used in the study was modeled after the AMPS-framework, these cues and their refined variations can make users reflective about their online sharing practices and help tackle problematic accounts.

Though we designed and studied the proposed intervention specifically for the retweeting scenario, the AMPS framework could be used more generally to help practitioners and researchers for devising a variety of cues for a wide range of decision making based on the background, history, and context of an account's activity, e.g., when deciding to follow/unfollow or friend/unfriend an account, deleting any past engagements with an account, etc. As platforms evolve, such a framework can promote the design of need-based contextual cues that help platforms to expose and (support user's agency to) counter sophisticated mechanisms of promoting problematic content.

7.4.2 Facilitate an easy access to online information trajectory. While approaches to help users investigate information propagation are effective [30], platforms need to incorporate them in user-friendly ways within everyday scenarios of use, e.g., retweeting on Twitter, forwarding on Whatsapp, sharing on Facebook etc. that need quick decision-making. One of the ways we realized this in our intervention was to only include three accounts: the root, a popularizer, and a friend. The design of trajectory that we studied led participants to assess their confidence in the propagating accounts and consequences of sharing that information further. We believe that different design choices—e.g., implementing the trajectory concept like an information accordion that gives on demand access to more accounts involved in the propagation—will lead to the discovery of more unique insights facilitated by information trajectories.

There is a reason to worry that providing access to such accounts and their activity might cause partisan sorting and increase online polarization [83]. We believe this might be compensated by accommodating instances where ideologically similar or familiar users share information with a diverse or dissenting perspective into the concept of trajectories. For example, including popularizers that used the quote-tweeting mechanism (on Twitter) or influencers that shared posts with a caption (on Instagram) to add an interesting but contradicting perspective. Participants from our study indicated that such accommodations within the design of trajectory might make them consider alternate perspectives.

7.4.3 Provide cues that vary as per the actor's role in information trajectory. We found that participants asked different questions when reflecting upon the different accounts involved in the trajectory. The need for unique credibility signals based on an account's role in dissipating an online post needs to translate into different informational cues that a platform offers its users. For example, providing cues about an account's expertise—especially for an account that is outside of one's curated network—only when that account plays a major role in spreading information. For within-network accounts, platforms can provide relationship-based signals, e.g., pop-up messages that question if they trust the user account, remind them that they recently followed/befriended the account from whom they are about to share content, etc. While more information on platforms will certainly add to the cognitive burden of processing it, providing relevant information in such a selective manner can assist online platforms to promote healthy discourse without limiting the user experience.

7.4.4 Highlight expert voices and institutions that popularize content. There is an urgent need for online platforms to focus on efforts that clearly signal what accounts brought attention to online content, i.e., popularize it. Following Hancock's suggestion [40, 63], such efforts will reduce distributed trust and help platforms refocus on institutional

trust by bringing user attention to the involvement (or its absence) of institutions like academia, medical authorities, etc. in popularizing information. For example, Twitter in 2020 took upon an initiative to verify the accounts of several medical professionals for identifying accurate COVID-related information [32].

7.4.5 Communicate information spread to users of variable information literacy. The information spread on social media is known to be networked. Though power users of these platforms (e.g., P4, P11) were aware of such networked-ness and its potential implications, more average users of these platforms could struggle to infer even the most basic signals by themselves. This limitation mandates platforms to devise ways in which they can communicate how information spreads to individuals that suit variable online information literacy, differing cognitive abilities to analyze and reflect upon information, and the *often* small attention span of decision making followed by users of the platform.

7.5 Limitations and Future Work

We designed our intervention to support the principles of the AMPS framework. Given the limitations of Twitter API, we could only access the 100 most recent retweets. As a result, we cannot say with certainty if the popularizer, whom we showed in the intervention, is a true popularizer for that tweet. Another limitation of the intervention is that we cannot confirm how or why a popularizer saw the tweet; it could be because they follow the root tweeter, or because it was promoted, or because one of their following accounts brought their attention to it. Similarly, we cannot confirm how and why the participant saw the tweet. Though the proposed intervention is not an accurate representation, it provides an approximation of how the tweet might have come across in the participant's feed. By ensuring a better connectivity of actors in the trajectory, there is an opportunity to investigate how knowledge about the nature of relationship/connection between two adjacent actors in the trajectory can impact content credibility.

We suspect that our findings might be impacted by some participant selection bias. Though we tried our best to recruit diverse participants in their life experiences and political ideologies, their willingness to participate in such a study can indicate their tendency to be critical about misinformation in general. We believe that users with a higher media literacy will potentially benefit more from such an intervention than those with a lower media literacy. Though reflective cues can be effective to promote critical thinking and discern hidden machinations of information online, automatically derived heuristic cues impose cognitive load on users when it comes to decision making [12]. To benefit from this realization, the current contextual cues can guide the future design of easy-to-use signals. For example, rather than showing the entire distribution of tweets, retweets, quote tweets and comments, use a simplified icon that conveys a relatively higher proportion of retweeting.

Our proposed choice—including the design of the intervention and the selected cues in it—is one of the many ways of designing such an intervention. It is possible that by changing the design of the intervention (colors of the cue card, the format of presenting the information, etc.) and offering different cues in the card might surface insights that diverge a little from the presented findings. We note that such a variation does not compromise the validity of the findings but is reflective of the research through design method [34] adopted in this research.

8 ACKNOWLEDGEMENTS

We thank University of Washington's Human-Centered Design and Engineering (HCDE) department and the Center For an Informed Public (CIP) for providing supportive communities. We also acknowledge the support of the National Science Foundation (NSF CAREER award no. 1749815) towards the facilitation of this research.

REFERENCES

[1] Swati Agarwal and Ashish Sureka. 2015. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*. Springer, 431–442.

[2] Jennifer Nancy Lee Allen, Cameron Martel, and David Rand. 2021. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. (2021).

[3] Ahmer Arif, John J Robinson, Stephanie A Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A closer look at the self-correcting crowd: Examining corrections in online rumors. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 155–168.

[4] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the part: Examining information operations within# BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.

[5] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[6] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.

[7] Namle Board. 2007. CORE PRINCIPLES OF MEDIA LITERACY EDUCATION in the UNITED STATES. (November 2007). https://namle.net/wp-content/uploads/2020/09/Namle-Core-Principles-of-MLE-in-the-United-States.pdf

[8] Jan Boehmer and Edson C Tandoc. 2015. Why we retweet: Factors influencing intentions to share sport news on Twitter. *International Journal of Sport Communication* 8, 2 (2015), 212–232.

[9] Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii international conference on system sciences*. IEEE, 1–10.

[10] Jessica E Brodsky, Patricia J Brooks, Donna Scimeca, Ralitsa Todorova, Peter Galati, Michael Batson, Robert Grosso, Michael Matthews, Victor Miller, and Michael Caulfield. 2021. Improving college students' fact-checking strategies through lateral reading instruction in a general education civics course. *Cognitive Research: Principles and Implications* 6, 1 (2021), 1–18.

[11] Monica Bulger and Patrick Davison. 2018. The promises, challenges, and futures of media literacy. *Journal of Media Literacy Education* 10, 1 (2018), 1–21.

[12] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.

[13] Fiona Carroll and Bastian Bonkel. 2021. Designing for affective warnings & cautions to protect against online misinformation threats. In *34th British HCI Conference 34*. 116–120.

[14] Fiona Carroll, Maggie Webb, and Simon Cropper. 2020. Investigating aesthetics to afford more 'felt'knowledge and 'meaningful'navigation interface designs. In *2020 24th International Conference Information Visualisation (IV)*. IEEE, 214–219.

[15] Der-Thanq Chen, Jing Wu, and Yu-Mei Wang. 2011. Unpacking new media literacy. (2011).

[16] Xusen Cheng, Shixuan Fu, and Gert-Jan de Vreede. 2017. Understanding trust influencing factors in social media communication: A qualitative study. *International Journal of Information Management* 37, 2 (2017), 25–35.

[17] Miyoung Chong and Hae Jung Maria Kim. 2020. Social roles and structural signatures of top influentials in the# prayforparis Twitter network. *Quality & Quantity* 54, 1 (2020), 315–333.

[18] Harris Cohen. 2021. Helpful Search tools for evaluating information online. *Google Blog* (September 2021). https://blog.google/products/search/evaluating-information-online-tools/

[19] Keith Coleman. 2021. Introducing Birdwatch, a community-based approach to misinformatio. (January 2021). https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

[20] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2451–2460.

[21] Nicholas Dias, Gordon Pennycook, and David G Rand. 2020. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review* 1, 1 (2020).

[22] Renee DiResta. 2021. The Misinformation Campaign Was Distinctly One-Sided. *The Atlantic* (March 2021). https://www.theatlantic.com/ideas/archive/2021/03/right-wing-propagandists-were-doing-something-unique/618267/?utm_source=twitter&utm_content=edit-promo&utm_term=2021-03-15T22%3A00%3A10&utm_campaign=the-atlantic&utm_medium=social

[23] Renee DiResta, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. 2019. The tactics & tropes of the Internet Research Agency. (2019).

[24] Judith Donath. 2007. Signals in social supernets. *Journal of Computer-Mediated Communication* 13, 1 (2007), 231–251.

[25] Steven Dow, Wendy Ju, and Wendy Mackay. 2013. Projection, Place and Point-of-view in Research through Design. *The SAGE Handbook of Digital Technology Research* (2013), 266–285.

[26] Stephanie Edgerly, Rachel R Mourão, Esther Thorson, and Samuel M Tham. 2020. When do audiences verify? How perceptions about message and source influence audience verification of news headlines. *Journalism & Mass Communication Quarterly* 97, 1 (2020), 52–71.

[27] Rosenberg Eli. 2019. How anonymous tweets helped ignite a national controversy over MAGA-hat teens. *The Hill* (2019). https://www.washingtonpost.com/technology/2019/01/23/how-anonymous-tweets-helped-ignite-national-controversy-over-maga-hat-teens/

[28] Miriam Fernandez and Harith Alani. 2018. Online misinformation: Challenges and future directions. In *Companion Proceedings of the The Web Conference 2018*. 595–602.

[29] Álvaro Figueira and Luciana Oliveira. 2017. The current state of fake news: challenges and opportunities. *Procedia Computer Science* 121 (2017), 817–825.

[30] Samantha Finn, Panagiotis Takis Metaxas, and Eni Mustafaraj. 2015. Spread and skepticism: Metrics of propagation on Twitter. In *Proceedings of the ACM Web Science Conference*. 1–2.

[31] Jennifer Fleming. 2014. Media literacy, news literacy, or news appreciation? A case study of the news literacy program at Stony Brook University. *Journalism & Mass Communication Educator* 69, 2 (2014), 146–165.

[32] Vijaya Gadde and Matt Derella. 2020. An update on our continuity strategy during COVID-19. (March 2020). https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19

[33] Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. 2016. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL* 24, 1 (2016), 42–53.

[34] William Gaver. 2012. What should we expect from research through design?. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 937–946.

[35] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[36] Nathaniel Gleicher. 2018. Coordinated Inauthentic Behavior Explained. *Facebook Blog* (December 2018). https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

[37] Sonja Grabner-Kräuter and Sofie Bitter. 2015. Trust in online social networks: A multifaceted perspective. In *Forum for social economics*, Vol. 44. Taylor & Francis, 48–68.

[38] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International conference on social informatics*. Springer, 228–243.

[39] Maria Haigh, Thomas Haigh, and Tetiana Matychak. 2019. Information literacy vs. fake news: the case of Ukraine. *Open Information Science* 3, 1 (2019), 154–165.

[40] Jeff Hancock. 2020. ET Speaker Series: Rethinking Trust and Well-Being in this Strange New World. (April 2020). https://www.youtube.com/watch?v=KLWCZNuopco

[41] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. 2013. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation* 4, 1 (2013), 3–28.

[42] Alfred Hermida, Fred Fletcher, Darryl Korell, and Donna Logan. 2012. Share, like, recommend: Decoding the social media news consumer. *Journalism studies* 13, 5-6 (2012), 815–824.

[43] Sally Rao Hill, Indrit Troshani, and Dezri Chandrasekar. 2017. Signalling effects of vlogger popularity on online consumers. *Journal of Computer Information Systems* (2017).

[44] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized social signals: Computationally-derived social signals from account histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[45] The Aspen Institute. 2021. Commission on Information Disorder: Final Report. *The Aspen Institute* (November 2021). https://www.aspeninstitute.org/wp-content/uploads/2021/11/Aspen-Institute_Commission-on-Information-Disorder_Final-Report.pdf

[46] Bahruz Jabiyev, Sinan Pehlivanoglu, Kaan Onarlioglu, and Engin Kirda. 2021. FADE: Detecting Fake News Articles on the Web. (2021).

[47] Caroline Jack. 2017. Lexicon of lies: Terms for problematic information. *Data & Society* 3, 22 (2017), 1094–1096.

[48] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–42.

[49] Minjeong Kang. 2010. Measuring social media credibility: A study on a measure of blog credibility. *Institute for Public Relations* (2010), 59–68.

[50] Douglas Kellner and Jeff Share. 2005. Media Literacy in the US. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung* 11 (2005), 1–21.

[51] Carolyn Mae Kim and William J Brown. 2015. Conceptualizing credibility in social media spaces of public relations. *Public Relations Journal* 9, 4 (2015), 1–17.

[52] James Klurfeld and Howard Schneider. 2014. News literacy: Teaching the internet generation to make reliable information choices. *Brookings Institution Research Paper* (2014).

[53] Timo Koch, Lena Frischlich, and Eva Lermer. 2021. The Effects of Warning Labels and Social Endorsement Cues on Credibility Perceptions of and Engagement Intentions with Fake News. (2021).

[54] Jiyoung Lee. 2020. The effect of web add-on correction and narrative correction on belief in misinformation depending on motivations for using social media. *Behaviour & Information Technology* (2020), 1–15.

[55] Stephan Lewandowsky and Sander Van Der Linden. 2021. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology* (2021), 1–38.

[56] Darren L Linvill and Patrick L Warren. 2020. Troll factories: Manufacturing specialized disinformation on Twitter. *Political Communication* 37, 4 (2020), 447–467.

[57] Andrew Lipsman, Graham Mudd, Mike Rich, and Sean Bruich. 2012. The power of "like": How brands reach (and influence) fans through social-media marketing. *Journal of Advertising research* 52, 1 (2012), 40–52.

[58] Sapna Maheshwari. 2016. How fake news goes viral: A case study. *The New York Times* 20 (2016).

[59] Zlatina Marinova, Jochen Spangenberg, Denis Teyssou, Symeon Papadopoulos, Nikos Sarris, Alexandre Alaphilippe, and Kalina Bontcheva. 2020. Weverify: Wider and enhanced verification for you project overview and tools. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–4.

[60] Diego A Martin and Jacob N Shapiro. 2019. Trends in online foreign influence efforts. *Princeton University, Princeton, NJ, Working Paper* (2019).

[61] Alice Marwick and Danah Boyd. 2011. To see and be seen: Celebrity practice on Twitter. *Convergence* 17, 2 (2011), 139–158.

[62] Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute* (2017), 7–19.

[63] Jesse McCrosky. 2020. How Social Media May Redistribute Trust Away From Institutions. (December 2020). https://dataethics.eu/how-social-media-may-redistribute-trust-away-from-institutions/

[64] Paul Mena. 2020. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & internet* 12, 2 (2020), 165–183.

[65] Finn S. Mustafaraj E. Metaxas, P. T. 2015. Using twittertrails. com to investigate rumor propagation. In *In Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work Social Computing*. 69–72.

[66] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication* 60, 3 (2010), 413–439.

[67] Maria D Molina, S Shyam Sundar, Thai Le, and Dongwon Lee. 2021. "Fake news" is not simply false information: a concept explication and taxonomy of online content. *American behavioral scientist* 65, 2 (2021), 180–212.

[68] Eni Mustafaraj and Panagiotis Takis Metaxas. 2017. The fake news spreading plague: was it preventable?. In *Proceedings of the 2017 ACM on web science conference*. 235–239.

[69] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5, 5 (2010), 411–419.

[70] Christina Peter and Thomas Koch. 2016. When debunking scientific myths fails (and when it does not) The backfire effect in the context of journalistic coverage and immediate judgments as prevention strategy. *Science Communication* 38, 1 (2016), 3–25.

[71] Nicolas Pröllochs. 2021. Community-Based Fact-Checking on Twitter's Birdwatch Platform. *arXiv preprint arXiv:2104.07175* (2021).

[72] Paul Resnick, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. 2014. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Proc. Computational Journalism Conference*, Vol. 5.

[73] Jon Roozenbeek and Sander van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5, 1 (2019), 1–10.

[74] Emily Saltz, Tommy Shane, Victoria Kwan, Claire Leibowicz, and Claire Wardle. 2020. It matters how platforms label manipulated media. Here are 12 principles designers should follow. *Partnership on AI* (2020).

[75] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*. 745–750.

[76] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592* 96 (2017), 104.

[77] Imani N Sherman, Jack W Stokes, and Elissa M Redmiles. 2021. Designing Media Provenance Indicators to Combat Fake Media. In *24th International Symposium on Research in Attacks, Intrusions and Defenses*. 324–339.

[78] Kate Starbird. 2020. Information operations and online activism within "NATO" discourse. *Three Tweets to Midnight: Effects of the Global Information Ecosystem on the Risk of Nuclear Conflict* (2020), 79–111.

[79] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[80] Brian Sternthal, Lynn W Phillips, and Ruby Dholakia. 1978. The persuasive effect of scarce credibility: a situational analysis. *Public Opinion Quarterly* 42, 3 (1978), 285–314.

[81] David Sterret, Dan Malato, Jennifer Benz, Liz Kantor, Trevor Tompson, Tom Rosenstiel, Jeff Sonderman, Kevin Loker, and Emily Swanson. 2018. *Who shared it? How Americans decide what news to trust on social media.* Technical Report. Norc Working Paper Series, WP-2018-001, 1–24.

[82] Briony Swire-Thompson, Joseph DeGutis, and David Lazer. 2020. Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition* (2020).

[83] Petter Törnberg. 2022. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences* 119, 42 (2022), e2207159119.

[84] Jason Turcotte, Chance York, Jacob Irving, Rosanne M Scholl, and Raymond J Pingree. 2015. News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of Computer-Mediated Communication* 20, 5 (2015), 520–535.

[85] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11.

[86] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[87] Emily K Vraga and Leticia Bode. 2021. Addressing COVID-19 misinformation on social media preemptively and responsively. *Emerging infectious diseases* 27, 2 (2021), 396.

[88] Emily K Vraga, Sojung Claire Kim, John Cook, and Leticia Bode. 2020. Testing the effectiveness of correction placement and type on Instagram. *The International Journal of Press/Politics* 25, 4 (2020), 632–652.

[89] Jen Weedon, William Nuland, and Alex Stamos. 2017. Information operations and Facebook. *Retrieved from Facebook: https://fbnewsroomus. files. wordpress. com/2017/04/facebook-and-information-operations-v1. pdf* (2017).

[90] David Westerman, Patric R Spence, and Brandon Van Der Heide. 2012. A social network as information: The effect of system generated reports of connectedness on credibility on Twitter. *Computers in Human Behavior* 28, 1 (2012), 199–206.

[91] Sam Wineburg and Sarah McGrew. 2017. Lateral reading: Reading less and learning more when evaluating digital information. (2017).

[92] Justine Wise. 2019. Twitter suspends account that helped incident with Native American man go viral. *The Hill* (2019). https://thehill.com/policy/technology/426338-twitter-suspends-account-that-helped-incident-involving-catholic-school

[93] Jingwen Zhang, Jieyu Ding Featherstone, Christopher Calabrese, and Magdalena Wojcieszak. 2021. Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines. *Preventive Medicine* 145 (2021), 106408.

[94] Yini Zhang, Josephine Lukito, Min-Hsin Su, Jiyoun Suk, Yiping Xia, Sang Jung Kim, Larissa Doroshenko, and Chris Wells. 2021. Assembling the networks and audiences of disinformation: How successful Russian IRA Twitter accounts built their followings, 2015–2017. *Journal of Communication* 71, 2 (2021), 305–331.

[95] Yuehua Zhao, Jingwei Da, and Jiaqi Yan. 2021. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Information Processing & Management* 58, 1 (2021), 102390.

[96] John Zimmerman and Jodi Forlizzi. 2014. Research through design in HCI. In *Ways of Knowing in HCI*. Springer, 167–189.

[97] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 493–502.