Investigating Information Provenance as a Cue to Look through the Opacity of (Mis)Information

Himanshu Zade

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2024

Reading Committee:
Jennifer Turns, Chair
Mark Zachry
Brock Craft
Ahmer Arif

Program Authorized to Offer Degree:
Human Centered Design and Engineering

University of Washington

Abstract

Investigating Information Provenance as a Cue to Look through the Opacity of (Mis)Information

Himanshu Zade

Chair of the Supervisory Committee:

Jennifer Turns
Human Centered Design and Engineering

Everyday users access information through different online media platforms as it propagates through the collective actions of users and algorithms. A diverse and huge audience with easy online access to such socio-technically curated information finds it challenging to make sense of this often evolving and sometimes contrasting information. Though researchers and some users know how to interpret this opacity of information, very few everyday users have the understanding and/or tools needed to look into online platforms' machinations and witness how and why (mis)information spreads. *How can we (re)design online media platforms to allow users to look through the opacity of propagating information — in often limited attention span of user interaction on these platforms — and assess the credibility of that information?*

This dissertation adopts information provenance — a record of information as it moves across users and platforms due to socio-technical actions — as a construct to ground thinking related to the overarching research question. Provenance facilitates a way to imagine what users could know about information and subsequently assess the credibility of the information.

The first study on the Google Search platform suggested the importance of information provenance. Findings suggest that making this provenance salient to users can convey the context behind information search and assist users in identifying how information surfacing through different provenances could vary in credibility. The second study on the Twitter platform utilized a design intervention that offered users a direct

window into parts of the provenance of information they were engaging with. Findings suggest that easy access to provenance can assist users in inferring specific judgments useful for credibility assessment. The third study on the TikTok platform revealed how users employ nuanced strategies unique to platform features to afford the socio-technical context of information. Findings suggest that when assessing credibility, users implicitly referenced the concept of provenance even without any direct provocation.

My dissertation presents numerous contributions across data, empirical research, artifacts, theory, and methodology based on studies conducted at three distinct research sites. Two of these contributions are particularly noteworthy. Firstly, I introduce a framework that proposes how design features of various platforms can facilitate credibility assessment. This framework considers dimensions such as inauthenticity, unfavorable online associations, contentious behavior, lack of trust, and the potential negative consequences of sharing information. Secondly, my dissertation establishes information provenance as a platform-agnostic cue for signaling information credibility. I discuss the findings in light of modern-day media literacy principles to advocate that information provenance could serve as a platform-agnostic construct valuable for conveying the opacity of information to users, who, in turn, can employ that realization to assess information credibility.

Acknowledgements

I want to extend heartfelt gratitude to my incredible advisors and mentors for their guidance and support that kept me on track throughout this wild research ride. A huge shoutout to my wonderful friends, family, and all the social connections who brightened my days and kept me sane. And of course, a special mention to those inspiring physical spaces where I conducted my research on online spaces — especially, those cozy cafes that fueled both my caffeine addiction and productivity. This dissertation wouldn't have been the same without you all!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Information that is rife with inaccuracy — intentionally or unintentionally (Jack, 2017) — seriously compromised the credibility of our information environments in the last decade. Propagation of such misinformation into our everyday news feeds can seriously erode the credibility of the information environments by creating confusion, chaos, and anxiety among the end users (Rapp and Salovich, 2018; Rocha et al., 2021). For successful discernment of misinformation, it now becomes essential for users to be able to assess its credibility by employing media literacy and for technology platforms to afford that credibility assessment by providing appropriate features.

Developing solutions that assist users in assessing the credibility of information in their news feeds remains a significant challenge, particularly due to perceptions that some solutions may be techno-centric (Friedman, 2019), struggle to keep pace with rapidly evolving adversaries, or even risk backfiring (Nyhan and Reifler, 2010; Schwarz et al., 2007). This challenge, along with the difficulty in defining the scope of the challenge, has led many researchers to characterize the spread of misinformation and problematic content online as a *wicked problem* (Jack, 2019; Wardle, 2018; Rittel and Webber, 1973).

Effectively addressing this complex issue necessitates a collaborative partnership between online media platforms, engineers, researchers, and policymakers. Such an alliance should involve jointly addressing both limitations and strengths, fostering continuous adaptation and evolution. However, challenges such as reductions in the technology labor force and concerns about regulating free speech have hindered the formation of such partnerships. Solutions that fail to facilitate this alliance may ultimately be inadequate in

addressing the scale of the challenge of assessing the credibility of information on online platforms.

Consider the example of the Covington Catholic High School controversy during the Indigenous Peoples March in Washington, D.C., in 2019. The confrontation that occurred between a group of High School students wearing "Make America Great Again" hats and Native American activist Nathan Phillips was reported in multiple versions online. As the messaging about who was at fault and who was the victim evolved, two information narratives grew prominent. Each version involved a video with misleading framing about the face-off and attributed responsibility for the confrontation to one of the involved parties. Given the limited context behind who is spreading the content, users were unable to assess credibility accurately. As a result, one of the videos got about 2.5 Million views and more than 14k retweets on Twitter platform (Wise, 2019); this 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 (Eli, 2019).

Information often gets reported using different narratives — be it due to the framing bias of a source or due to evolving details about the incident — making it challenging to assess its credibility. Modern-day online platforms, however, offer *technological* affordances that could rapidly amplify the reach of these narratives through the agency of *social* users, as seen in the case of the Covington High School controversy mentioned above. A diverse and massive audience with easy online access to these shifting and often contrasting narratives finds it challenging to engage with content in such an environment as they struggle when sensemaking (Sandberg and Tsoukas, 2015) of the narratives and making judgments about which narrative to believe. This condition is only worsened by any political *socio-technical imaginaries* to intentionally propagate contrasting or state-promoted information narratives online (Bächtold, 2023). Despite this challenge of sensemaking and discerning credible information, platform designs offer little to help users understand the machinations of this information environment — also referred to as *information fog* (Badke, 2021) — caused by these relatively new and emerging socio-technical affordances, thereby exacerbating the challenge of assessing information credibility. I refer to this condition of insufficient transparency regarding the dissemination of information online before it reaches a user's online feed as *opacity*. I elaborate more about the term 'opacity' in Chapter 2. For humans to meaningfully and ethically engage with an increasingly complex information environment, Vallor insisted that humans need to learn how to deal with this opacity (Vallor,

2016):

> "Techno-moral practices on a global scale remain extremely immature from a developmental perspective as a result of how recent the new technological affordances for global action have emerged." - Shannon Vallor (pg. 48)

How can we (re)design online media platforms and impart in users the ability to look through the information opacity — in often limited attention span of user interaction on these platforms — and understand why problematic information spreads? To answer this question and understand how misinformation operates, it is valuable to look at multiple social media platforms, given their unique affordances and challenges. The dissertation discusses three studies on Google Search, Twitter, and TikTok platforms and demonstrates how to witness, interpret, and signal the opacity of information. The findings suggest lessons that will help everyday users to engage with modern media platforms meaningfully.

Misinformation researchers and journalists have conducted analyses on how misinformation spreads to protect the public from problematic content. Much of their understanding of misinformation propagation comes from studying the broader network of online actors involved, including influencers, political figures, and information organizations (Liang et al., 2019; Stewart et al., 2018; Schafer and Starbird, 2023). Additionally, regular users sometimes contribute to this propagation by sharing information without effectively assessing its credibility or understanding the reasons behind its spread. The documented path that traces the origin of information and its movement from user to user, and sometimes across platforms, is referred to as its provenance. To assess the credibility of online information and determine whether it has the potential to misinform, it is crucial to obtain additional context about its provenance.

Researchers have studied the impact of providing additional context in regards to algorithms and explainable artificial intelligence (AI); *e.g.*, Eslami found that helping users interpret the opacity of algorithmic decisions in online spaces can improve users' understandability of automated decisions, thereby fostering a more informed and satisfying user engagement (Eslamimehdiabadi, 2019). This dissertation extends scholarship around the construct of opacity by focusing on information provenance and helping users assess the credibility of information as it circulates online.

Researchers have developed mechanisms based on the diffusion of information to help users assess credibility and identify problematic content (Marinova et al., 2020; Resnick et al., 2014). For instance, Finn

et al. created a tool called *Twitter trails* to aid in investigating a tweet's potential as a rumor based on its propagation within a social network (Finn et al., 2015; Metaxas, 2015). However, these mechanisms are tailored to the investigative needs of researchers or journalists and may not meet the needs of everyday social media users. Therefore, I focus on designing platforms that assist users in engaging with the opacity of information and learning about its origins before interacting with it.

Search engines often serve as the initial gateway for users to access information. One might assume that the online record of information maintains its credibility uniformly regardless of the medium or location of access. I tested this hypothesis by examining headlines on the Google search platform. As information spreads, users contribute rich context and interpretations, leading to increasingly complex trajectories of related information. Some of these trajectories may be rapidly amplified for inappropriate reasons, forming the core of my second investigation on the Twitter platform. Lessons learned about avoiding misinformation on Twitter may not apply to platforms like TikTok due to their distinct features and largely hidden algorithmic mediation, which is more prevalent on TikTok. Therefore, I identified how one can generalize the learning about assessing information credibility beyond Twitter, which forms the basis of my third investigation.

This dissertation examines online information platforms to witness, interpret, and signal the opacity of information with a particular focus on information provenance. It investigates how information provenance could potentially serve as a platform-agnostic construct to convey the opacity of information to users, who, in turn, can employ that realization to assess the credibility of that information. It introduces three studies conducted at three unique sites of research, Google Search, Twitter, and TikTok, to examine the potential of designing infrastructures that help everyday users understand information propagation and assess the validity of their everyday online feeds. Contributions include:

- **Witnessing information opacity on Google Search:** Given the prominence of search engine platforms in everyday lives to get credible news, these platforms are important gateways for diverse users to access information through different media. *RQ1:* As information surfaces within a platform (Google Search), how do different contextual factors behind that information impact its credibility?

  I conducted a content analysis of headlines appearing on Google search engine result pages (SERPs) in response to 20 election-related keywords — 10 general (e.g., 'Ballots') and 10 conspiratorial (e.g., 'Voter fraud') — when searched from 20 cities across 18 states. Results revealed the asymmetry

of how some information provenances consisting of videos (compared to stories, search results, and advertisements) can contribute more delegitimizing content.

- **Interpreting information opacity on Twitter:** Users of online platforms play an active role in spreading the evolving information and its interpretations as it propagates. *RQ2:* How can easy access to informational context — presented as a provenance — on a platform (Twitter) help users assess the credibility of that information?

  For conceptualizing provenance, I introduce the 'Twitter trajectory' intervention to provide informational context — consisting of an information trajectory and account activity cues — and demonstrate, using research through design, how it can equip users to identify problematic behaviors (inauthenticity, unfavorable online associations, contentious behavior, lack of trust, unwanted consequences of sharing) of online accounts that one must learn to assess the account's credibility at a broader level — that of the overall network involved in spreading information — and not merely at the level of an individual account. The proposed intervention and its variations offer a foundation for future researchers to ask many interesting questions and discover insights about facilitating credibility assessment within propagation networks.

- **Signaling information opacity on TikTok:** Though I identified different lessons that can help understand how misinformation propagates on a platform like Twitter, these lessons may translate poorly to other platforms. *RQ3:* When moving to platforms beyond Twitter (TikTok), how do users employ the platform cues to identify the socio-technical context of information on that platform?

  I conducted participatory design research and analyzed self-reported TikTok users' experiences to identify what design features of the TikTok platform are useful to signal socio-technical dimensions of video-based information on TikTok. The unique features of TikTok (*e.g.*, stitching other creators' content to add one's perspective) afforded distinct user engagement as compared to Twitter but afforded similar processes when it came to judging information credibility.

Collectively, the dissertation offers a cohesive picture of how platforms can offer everyday users a well-researched and contextualized approach about *where to look* and *what to know* when assessing the credibility of information. As I discuss the findings, I also elaborate on how the cues that signal the opacity of infor-

mation can be generalized across different platforms — whether more user-driven or algorithm-driven — and support users towards safe discernment of that (mis)information. I promote making these cues and the associated learning a part of essential media literacy across different platforms.

Chapter 2

Background

In this chapter, I begin by outlining the socio-technical complexities involved in assessing the credibility of online information. I discuss various forms of problematic content, delve into the socio-technical aspects of information, and introduce the concept of *opacity* to characterize the contextual challenges. Next, I review influential research on credibility assessment, highlighting different social media cues that aid in evaluating information credibility both independently and within its propagation network. Third, I delve into the notion of media literacy-based initiatives and how they have evolved over the years. Finally, I discuss ways to implement these media literacy-based strategies effectively on emerging media platforms like Snapchat, WeChat, and Instagram to assist credibility assessment efforts better and bolster platforms' resilience against misinformation.

2.1 The challenge of credibility assessment

2.1.1 Types of problematic information

The information we see online can sometimes be compromised regarding its credibility. False information that spreads online can have some serious consequences. For example, election-related misleading content that suggests irregularities in the elections can cause serious damage to the public trust in media and institutions (Ognyanova et al., 2020; Lupu et al., 2020). The spread of such misleading content into our media feeds compromises our collective ability to establish a common ground about topics of civic interest.

To best support efforts in tackling problematic content, it is important to establish a common vocabulary for facilitating collective effort towards tackling such content. For establishing a taxonomy of different kinds of problematic content, Molina et al. characterized news articles that we might commonly refer to as *fake news* into seven different types: false news, polarized content, satire, misreporting, commentary, persuasive information, and citizen journalism that based on different features evident within that news article (Molina et al., 2021). Their classification effort utilized the factual accuracy of the information in the article as one of the important features. They suggested that only the *false news* category was distinctly non-factual, whereas the other six categories varied along the spectrum between non-factual and factual. In the scope of this research, all the articles along this spectrum between non-factual and factual could be seen as misinformation. Focusing on another feature *consensus*, Southwell et al. defined misinformation as "claims that do not enjoy universal or near-universal consensus as being true at a particular moment in time on the basis of evidence" (Southwell et al., 2017). In another body of research, 'Lexicon of Lies,' Jack focused on the *intention* behind the article's potential inaccuracy and described, "Misinformation is information whose inaccuracy is unintentional. Disinformation is information that is deliberately false or misleading" (Jack, 2017).

Regardless of the type of problematic content, platforms need to design solutions that assist users in assessing the credibility of information as users struggle to decide if what they see is credible or not. In this dissertation, I do not focus on any specific type of problematic content but instead focus on the efforts that assist in assessing the credibility of information.

2.1.2 Socio-technical characteristics of online (mis)information

Online platforms like Facebook, Twitter, etc. have made it easier to spread misinformation or to engage in inauthentic online behavior (Jack, 2017; Martin and Shapiro, 2019; Starbird et al., 2019; Weedon et al., 2017; Arif et al., 2018). 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 (Linvill and Warren, 2020; Zhang et al., 2021b; DiResta et al., 2019), how internet groups manipulated and propagated selective news frames that initially surfaced on *8chan* (Marwick and Lewis, 2017), etc. Another example is the political *socio-technical imaginaries* of Myanmar engaging in

intentional propagation of state-promoted misinformation online (Bächtold, 2023).

In order to carefully navigate the potentially problematic information online, it is important for users of the media platforms to armor themselves with knowledge and skills to identify such inauthentic organized efforts (Gleicher, 2018). Despite the challenges posed by this socio-technically orchestrated problematic content, the design of media platforms offers little to support users in identifying and making sense of these information machinations, including inauthentic organized efforts (Gleicher, 2018; Starbird et al., 2019), false amplification of content using accounts with fake online profiles (Mazza et al., 2022; Kwon et al., 2022), popularizing government propaganda (Chang et al., 2021; Zannettou et al., 2019), etc. The complexity behind the machinations and limited platform support for users to recognize these machinations impart in the online spaces a sense of opacity that I discuss next.

Information Opacity

Humans have often employed technological affordances to create complex socio-technical conditions that they are ill-prepared to deal with fully. The challenge of assessing information credibility online to decide if something is misleading or not is one such condition. Addressing this challenge would be more manageable if users had convenient access to the socio-technical context surrounding the information. However, current platform designs provide minimal assistance in elucidating the inner workings of information ecosystems, rendering valuable characteristics opaque to users.

Badke invoked the term *information fog* to refer to the overwhelming volume of information available to researchers and how it challenges navigating and evaluating the modern information landscape effectively (Badke, 2021). For humans to meaningfully and ethically engage with the increasingly complex information environment, Vallor insisted that humans need to learn how to deal with this sense of opacity that is socio-technical in nature (Vallor, 2016).

Researchers have explored the concept of opacity in the realm of algorithms, aiming to enhance the explainability of these systems for users. For instance, efforts have focused on improving users' comprehension of artificial intelligence (AI) and enabling constructive engagement with it in various aspects of human life (Hase and Bansal, 2020; Shin, 2021). In a related study, Eslami demonstrated that aiding users in interpreting the socio-technical opacity surrounding algorithmic decisions in online environments can en-

hance their understanding of automated decisions, leading to more informed and satisfying user interactions with AI systems (Eslamimehdiabadi, 2019). Concerns about the opacity of interpreting AI systems have also been raised due to ethical issues such as bias and discrimination (Crawford and Calo, 2016). Furthermore, in the context of privacy, Hargittai et al. used the term opacity to describe users' challenges in understanding online privacy settings due to their complexity (Hargittai, 2007). Indeed, research across disciplines has shown that increasing the transparency of systems can enhance users' trust in those systems (Peeters and Pulls, 2015; Grimmelikhuijsen et al., 2013).

I embrace the term *opacity* to align my research with existing literature exploring the advantages of enhancing transparency to foster credibility. Within this dissertation, I define opacity as the condition of insufficient transparency regarding the dissemination of information online before it reaches a user's online feed. Expanding on this notion of opacity, I introduce information provenance — *i.e.*, the documentation of how information circulates among users and platforms — to aid users in evaluating the credibility of propagating information.

2.2 Signals for assessing information credibility

The widespread use of digital media has provided users on online platforms with access to a vast amount of information. However, determining which information to trust and engage with is increasingly challenging. Therefore, evaluating the credibility of the information encountered online is paramount. While credibility typically encompasses dimensions of trust and expertise (Hovland et al., 1953), the process of assessing credibility focuses on determining the believability of a source or message (Metzger and Flanagin, 2015). The approach to assessing credibility varies across disciplines; for example, information science researchers prioritize the reliability and accuracy of data and sources, while communication researchers consider credibility as a perceived characteristic influenced by the message, source, and media (Rieh and Danielson, 2007). Given the lack of a universally agreed-upon perspective on credibility (Hilligoss and Rieh, 2008), this dissertation adopts "a blend of subjective and objective assessments" (Bhuiyan et al., 2020) to accommodate the interdisciplinary nature of the research presented herein.

A combination of subjective and objective evaluations is crucial for assessing credibility, aiding in the identification of nuanced attempts to deceive. For instance, a 2011 study highlighted that features related

to the topic, user, and propagation, such as a higher number of retweets, were considered most suitable for shaping credibility judgments on the Twitter platform (Castillo et al., 2011). However, with the rise of information warfare, merely relying on the number of retweets without considering the subjective context of the retweeter is insufficient for effective credibility assessment in today's online social networks. Research suggests that broadening the scope beyond a checklist of features to encompass the wider social context of information can lead to more informed judgments about information quality (Meola, 2004). For instance, increased monitoring of content by moderators has been shown to enhance the overall credibility of online communities (Hajli et al., 2015). To enhance user motivation for assessing credibility (Metzger, 2007), this dissertation advocates for making contextual factors readily accessible through provenance-based design, thus aiding in the investigation of information opacity. The subsequent section introduces how the signals facilitating credibility assessment online have evolved over time.

2.2.1 Evolution of social media signals for assessing credibility

During the late 2000s, social media platforms emphasized popularity metrics such as the count of likes and shares. These metrics served as indicators for guiding users on how to interact with online content (Hermida et al., 2012; Lipsman et al., 2012; Hill et al., 2017). For curtailing the spread of problematic content to create healthier online spaces, the design of these platforms evolved to include warning and corrective labels that provide additional context as considered appropriate by the platforms (Zhang et al., 2021a; Mena, 2020; Lee, 2020; Vraga et al., 2020; Koch et al., 2021). Preemptive labels aimed at inoculation have been effective in helping people assess credibility by warning them of how they can be misinformed, thereby increasing their resilience to misinformation (Lewandowsky and Van Der Linden, 2021; Vraga and Bode, 2021; Roozenbeek and van der Linden, 2019). Corrective labels, while usually helpful, 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 (Bail et al., 2018; Peter and Koch, 2016; Swire-Thompson et al., 2020).

To support regular platform users and guide the application of these labels, in 2021, Twitter introduced a community-based approach called Birdwatch (Coleman, 2021; Pröllochs, 2022) so that users can add context to a tweet. Such labeling, unfortunately, has been found to support partisanship rather than promote fact-checking (Allen et al., 2022). Automated algorithms trained on human-labeled data about tweets' cred-

ibility have sometimes rendered users unhappy as they disagreed with the credibility labels due to individual differences of what and whom they considered credible (Gupta et al., 2014).

The ongoing effort to incorporate automatically generated credibility signals has been successful to a significant extent (Kang, 2010; Donath, 2007; Gupta et al., 2014). This is because automatically generated credibility signals are challenging for users to manipulate compared to more traditional signals like their profile information. For example, using labels that reflect the accuracy of headlines has been found effective in reducing the sharing of misinformation (Jahanbakhsh et al., 2021). Researchers have demonstrated that automatically derived nudges can be effective in providing some context into the credibility of information (Im et al., 2020; Bhuiyan et al., 2021). Unfortunately, the labels implemented by platforms have not consistently proven effective; users tend to be more inclined to consider verification suggestions provided by these prompts when they perceive that the message aligns with their own ideologies (Edgerly et al., 2020). In addition, these approaches are more suitable to the automatic way of thinking (over reflective) (Caraban et al., 2019; Hansen and Jespersen, 2013) to support quick decision-making. To help identify information machinations of a more organized nature and preserve user agency, this dissertation promotes reflective strategies to be more effective and resilient as platforms, and the underlying challenges evolve continuously.

2.2.2 Nuanced social media signals based on information propagation and provenance

Researchers are aware that information spreads online through a network (Arif et al., 2018; Starbird, 2020). Network properties and behavioral features of propagation help discern misinformation from good information (Zhao et al., 2021; Molina et al., 2021); e.g., witnessing superficial characteristics of a Twitter account's social network (e.g., information about followers and following) can assist users in making an informed decision about sharing content from that account (Westerman et al., 2012).

Misinformation researchers and journalists have conducted their own analysis about misinformation propagation to protect the public from problematic content. Although what is considered problematic varies on the context — *e.g.*, content that is corrected by fact-checking websites when doing behavioral research on misinformation (Pennycook et al., 2021), content that is deceptive relative to the best available scientific evidence in case of identifying scientific misinformation (Southwell et al., 2022), content that may erode public trust in civic institutions in case of elections (Zade et al., 2022b), — the learning about how that prob-

lematic content spreads primarily comes from understanding the broader network of online actors involved in propagating misinformation. These actors might include influencers, political actors, information organizations (Liang et al., 2019; Stewart et al., 2018; Schafer and Starbird, 2023), and most importantly, regular users who sometimes partake in such propagation when they share posts without effectively assessing the credibility of sources or the reasons behind the spread of information. The documented trail that accounts for the origin of information and how it moves around from user to user and, at times, from platform to platform can be referred to as its provenance. To assess the credibility of online content and discern if the information is misleading, it becomes essential to get additional context about its provenance — and to design social signals that capture and communicate this context.

The original source of information is considered one of the prime factors for assessing the credibility of content (Metzger et al., 2010; Jabiyev et al., 2021). Despite the challenges of identifying the information source and verifying its credibility for researchers and media platforms (Starbird et al., 2019; Diakopoulos et al., 2012; Figueira and Oliveira, 2017), cues about the source of information may not always be effective towards assessing content-credibility and detecting misinformation (Dias et al., 2020).

To benefit from other signals apart from the source, researchers have utilized diffusion-based metrics to help users assess the credibility of the content (Marinova et al., 2020; Resnick et al., 2014). For example, Finn et al. developed a tool called *Twitter trails* to help users investigate a tweet of their interest if it is a potential rumor based on how it propagates within a social network (Finn et al., 2015; Metaxas, 2015). Shao introduced the Hoaxy platform along similar lines to facilitate the collection, detection, and analysis of all incoming tweets to detect misinformation online (Shao et al., 2016). While these platforms are extremely useful for assessing credibility, they have been designed in a way that is more suitable for an investigative perspective like that of a journalist or researcher than that of an average social media user.

2.3 Incorporating media literacy initiatives when assessing credibility

Media literacy is another necessary discipline that has contributed immensely to helping users live well in online information environments like Facebook, Twitter, Google Search, and others. With the rapidly evolving challenge posed by misinformation, there is a shared need for platforms and their users to critically inspect how media literacy principles can be employed in our everyday lives to learn about the opacity of

information propagation.

While disinformation distracts and confuses the intended audience using deliberate and misleading information, misinformation could be unintentionally inaccurate (Wu et al., 2019). When an abundance of information is floating around (*e.g.*, breaking news), uncertain information and rumors might be floating around that are later corrected by reputed and credible news sources. However, platform users may already have been exposed to misinformation by that point. Therefore, platform users must discern between uncertain information and what is undoubtedly misinformation. This becomes particularly critical as users are less likely to question information when it comes to them through a network that they trust, thus lowering their attention vigilance (Pennycook and Rand, 2021b; Sterret et al., 2018). In a situation of high criticality, exposure to any misleading information can sometimes form the basis of why we vote and how we vote (Diakopoulos et al., 2018)! In order to carefully navigate potentially polarizing and sometimes manufactured information online, users need to arm themselves with knowledge and skills to identify such inauthentic organized efforts as misinformation (Gleicher, 2018).

Media literacy became popular in the early 1960s to empower students to use media, namely newspapers and television, to consume information responsibly. McLuhan, through his book 'The Medium Is the Message,' was an influential force behind this move and suggested "the personal and social consequences of any medium depend on the scale introduced by any new tech into our affairs of the world" (McLuhan and Fiore, 1967). In other words, how information is delivered changes our relationship with that information. With the recent adoption of new media platforms in our everyday lives and the emergence of computational propaganda (Woolley and Howard, 2016; Arnaudo, 2017; Woolley and Howard, 2017; DiResta, 2018), *i.e.*, using computer-controlled accounts to influence information propagation and amplifying their reach strategically, it has become even more critical to adopt media literacy in practice to be able to analyze a media message (Delwiche and Herring, 2020). Before I discuss more about such adoption practices, let us dig deeper into what literacy and media literacy constitute.

2.3.1 Understanding literacy

Until 1966, the concept of literacy was often understood to be an individual's ability to read and write. In 1966, the United Nations Educational, Scientific and Cultural Organization (UNESCO) set up the World

Literacy Program and moved from the skill-based definition of literacy to a more functional one (Spaulding, 1966). The new functionality-based definition identified a person as literate provided "they can engage in all the activities in which literacy is required for them to effectively function in their group and community and also for enabling them to continue to use reading, writing, and calculation for their own and their community's development." This shift from a skill-based to a functionality-based literacy definition was also adopted by the American National Assessment of Adult Literacy (NAAL), much later in 2003, as they updated their view of literacy from "using printed material successfully" to "making use of printed material to function in the society and achieve collective benefit" (White and McCloskey, 2003).

In a seminal work on 'The Right to Literacy', Knoblauch suggested another shift in the perspective of what constitutes literacy from thinking about it as an objective capacity to a subjective judgment (Knoblauch, 1990). Knoblauch offered three definitions of literacy that consider an individual's culture, personal growth, and critical thinking. Each of these perspectives has some drawbacks — *e.g.*, while a functional perspective can be ableist, the cultural perspective can lead to the marginalization of non-dominant groups — but all these perspectives assert that literacy is an individual's responsibility.

Overcoming the limitations of an individualistic definition of literacy that considers it in isolation from an individual's surroundings and social experiences, some scholars like Scribner (1984) suggested thinking of literacy as a more collective responsibility and referred to literacy as a social achievement given that it can only be attained by participating in socially organized activities (Scribner, 1984). Kliewer found support for this view through ethnographic activities studying sensemaking in disabled children and suggested that we define literacy as the construction — including collective interpretation — of meaning through visually or tactually crafted symbols that compose various forms of text (Kliewer et al., 2004; Kliewer, 2008).

These shifting perspectives from ability-based to functional, objective to subjective, and individualistic to collective are essential to attain literacy that is suitable for critical thinking. The different kinds of literacies — technological, digital, computer-related, informational, library, network, and media-related (Bawden, 2001) — are all suited for different contexts and informational requirements (Bawden et al., 2008). Out of these literacies, media literacy, *i.e.*, one's ability to use that information upon granted access through any media is more important to evaluate the message critically (Koltay, 2011). From here on, with a focus on facilitating such critical evaluation, I dig deeper into media literacy.

2.3.2 Understanding media literacy

Just like literacy, what constitutes media literacy has evolved to accommodate the changing nature of media forms, including Snapchat and Instagram stories today, or perhaps augmented reality (AR) and interactive scenarios in the very near future. As each new communication medium presents a unique challenge, humans are compelled to adapt and cultivate literacy in these emerging forms of media. Academics and research organizations have adapted their definition of literacy, as has the US government and the UK government. According to a cultural commentator:

> "Media literacy is nothing new, but it's adapting and changing all the time. Where media literacy once required a mastery of language and a quill, the age of the penny press required the ability to analyze headlines at a glance and tell the truth from sensationalism." - Jay Smooth (Smooth, 2018)

As scholars address media, literacy, and the purpose of media literacy across several suggested definitions (Potter, 2013), three things stand out. First, scholars believe that media literacy involves decoding the hidden meaning or intent behind information in the media. Second, they believe such an interpretation of that information is subject to an individual's experiences and, hence, deeply rooted in that interpretation's sociocultural, socio-economic, and socio-political context. Third, media literacy doesn't stop once we infer the nuances within that information; it also includes how we incorporate that meaning as we communicate further through more media messaging. This dissertation looks at the implications of these steps as they relate to assessing the credibility of information and maybe sharing information with other users on the platform.

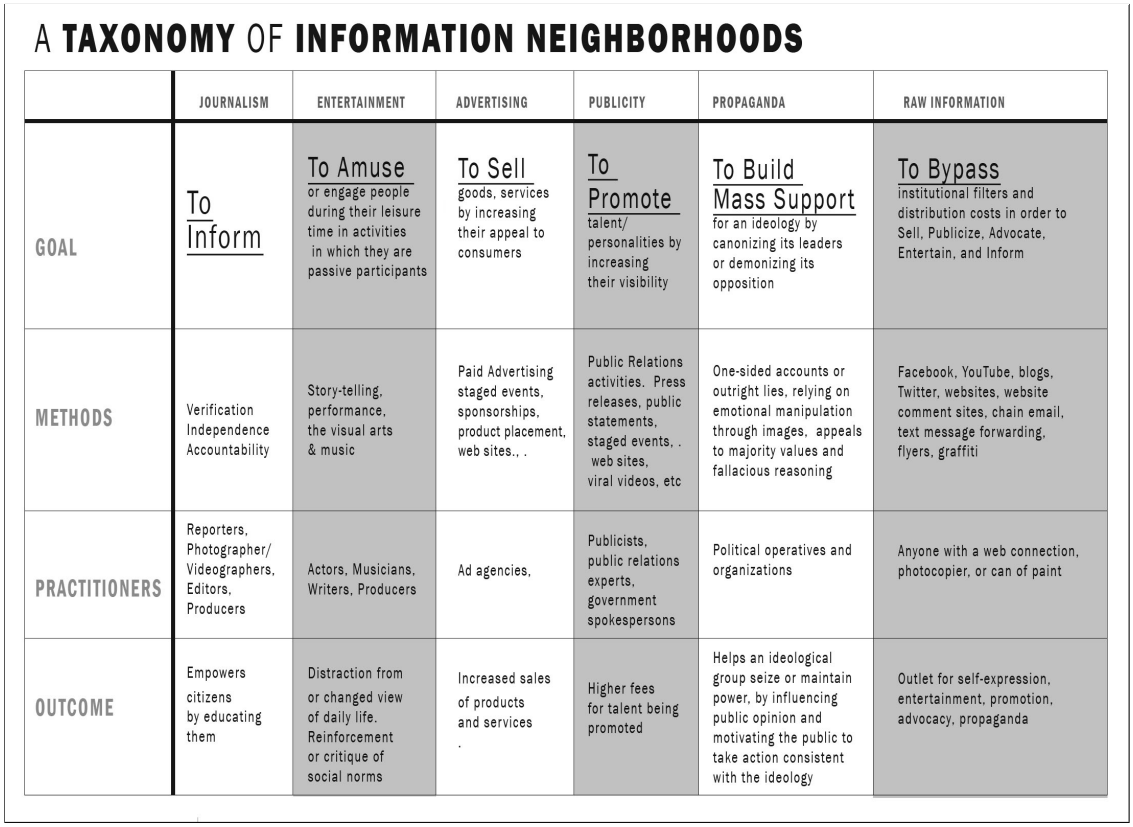
To capture these different components of media literacy as they relate to the evolving nature of media, I adopt the definition of digital media literacy as the ability to access, analyze, evaluate, create, and act using all forms of information (Board, 2007). The first component of media literacy *access* refers to the ability to get hold of information, which includes both media (devices and services) and the actual content published through those media. Once a user has access to the information of interest, it is important that they can decode that information and *analyze* it. The next step then involves interpreting that derived meaning to *evaluate* it as per one's evaluation process and personal experiences. The last two parts constitute

thoughtfully *creating* new information and responsibly *acting* upon any information.

2.3.3 Teaching media literacy

Media literacy scholars have suggested different ways of incorporating digital media literacy principles — access, analyze, evaluate, create, and act — into media-based learning. One of the earliest ones included the Canadian communication scholar Marshall McLuhan, who ignited the North American educational movement for media literacy in the 1950s and 1960s (McLuhan and Fiore, 1967). He coined the phrase 'media is the message' to refer to the symbiotic relationship between a medium and a message, inspiring media literacy research for decades. Renee Hobbs is another internationally recognized digital and media literacy education authority affiliated with communication studies at the University of Rhode Island's Harrington School of Communication and Media (Hobbs, 2021). In her book 'Media Literacy in Action: Questioning the Media,' she suggests a reader ask five questions when evaluating the accuracy of information about the purpose, attention-attracting techniques, value system behind the shared perspective, possible interpretations of that information, and omitted details. Another perspective involves facilitating critical comprehension of information for discerning good media from the bad and accounting for a creator's accuracy, bias, and reliability based on the creator's history (Coiro, 2014). Stony Brook University's model of teaching media literacy that focuses on using a taxonomy of information neighborhood (as shown in Figure 2.1) has been found effective in guiding more than ten thousand young minds into thinking if the information they see fits the standard of journalistic information or if it seems propaganda-like (Fleming, 2014).

The developments of the media literacy curricula are widespread even outside of the USA (though less pronounced). For example, Kajimoto identified the common patterns of news consumption across three Asian countries (Hong Kong, Vietnam, and Myanmar) and suggested how the region-specific issues need to be integrated into the development of news literacy curricula suitable to the local circumstances (Kajimoto, 2016). In another related attempt, educators tried to come up with local examples by replacing the American references with more relatable Chinese references to help non-American countries learn and benefit from the lessons on media education and develop media-related competency, *e.g.*, the inclusion of local news reports about an unannounced concert by Psy of *Gangnam Style* fame (Hornik and Kajimoto, 2014). To support users in being critical of the information they consume, in 2015, Ukraine administered the 'Learn to Discern'



A **TAXONOMY** OF **INFORMATION NEIGHBORHOODS**

	JOURNALISM	ENTERTAINMENT	ADVERTISING	PUBLICITY	PROPAGANDA	RAW INFORMATION
GOAL	To Inform	To Amuse or engage people during their leisure time in activities in which they are passive participants	To Sell goods, services by increasing their appeal to consumers	To Promote talent/ personalities by increasing their visibility	To Build Mass Support for an ideology by canonizing its leaders or demonizing its opposition	To Bypass institutional filters and distribution costs in order to Sell, Publicize, Advocate, Entertain, and Inform
METHODS	Verification Independence Accountability	Story-telling, performance, the visual arts & music	Paid Advertising staged events, sponsorships, product placement, web sites., .	Public Relations activities. Press releases, public statements, staged events, . web sites, viral videos, etc	One-sided accounts or outright lies, relying on emotional manipulation through images, appeals to majority values and fallacious reasoning	Facebook, YouTube, blogs, Twitter, websites, website comment sites, chain email, text message forwarding, flyers, graffiti
PRACTITIONERS	Reporters, Photographer/ Videographers, Editors, Producers	Actors, Musicians, Writers, Producers	Ad agencies,	Publicists, public relations experts, government spokespersons	Political operatives and organizations	Anyone with a web connection, photocopier, or can of paint
OUTCOME	Empowers citizens by educating them	Distraction from or changed view of daily life. Reinforcement or critique of social norms	Increased sales of products and services .	Higher fees for talent being promoted	Helps an ideological group seize or maintain power, by influencing public opinion and motivating the public to take action consistent with the ideology	Outlet for self-expression, entertainment, promotion, advocacy, propaganda

Figure 2.1: Using a Taxonomy of Information Neighborhoods, students at Stony Brook University's Journalism school learn how to characterize information in a systematic way to distinguish characteristics of news, promotion, propaganda, and raw information.

information literacy program situated in the local context as a response to the flood of misinformation produced by Russia in 2014 (Haigh et al., 2019). The program was also effective in helping users discern the reliability of sources when presented with conflicting information.

2.4 Media literacy-based approach to reduce information opacity when assessing credibility

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 (Wu and Wang, 2011). 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 (Board, 2007; Kellner and Share, 2005). Unfortunately, such media literacy efforts largely have been understood and promoted as an individual's responsibility. Such a singular perspective on media literacy is inadequate to identify modern-day information machinations. To develop media literacy efforts that help users look through the opacity of online information, platforms, and their users need to realize the limitations of disconnected individual experiences and define and implement them at the level of an organization, platform, and nation (Bulger and Davison, 2018).

> "Misinformation is not like a plumbing problem you fix. It is a social condition, like crime, that you must constantly monitor and adjust to." - Tom Rosenstiel (Anderson and Rainie, 2017)

As the persuasiveness of media technologies has catalyzed McLuhan's conceptualization of the *global village* where a diverse and massive audience gets quick access to sometimes varying information about an event, it is critical to evaluate that information by questioning its whereabouts and negotiating the boundary between what could be the true meaning and what could be a possible misrepresentation (McLuhan and Powers, 1989). This notion of *global village* reappears in Vallor's writing, many decades later in 2017, as she discusses the construct of opacity so users of diverse cultural backgrounds can not only access but also meaningfully engage in the plurality of interpretations without getting confused by the contrasting information (Vallor, 2016).

For facilitating platforms that are always ready to tackle new kinds and variations of problematic content, it is, therefore, essential to impart a sense of literacy into them. By pushing the onus of learning literacy away from that of users towards the platforms, newer media platforms can be built with a focus on design features rooted in critical thinking. Susceptibility to sharing misinformation has been accounted more to lack of attention than ill intention (Ross et al., 2021). Therefore, I see tremendous hope in the media literacy-based approach that I discuss in this dissertation to employ provenance-centered solutions to educate users about the opacity of information and help users become more aware of how and why information propagates as they assess the credibility of that information.

Chapter 3

Methodology

In this chapter, I introduce the overall methodology adopted for the research presented in this dissertation. I followed a (mostly) qualitative approach consisting of audit-based content analysis, research through design, and participatory design techniques to answer different questions of interest about assessing the credibility of information by harnessing the socio-technical context in which information propagates across different media platforms.

I break down the primary question of *how can we (re)design online media platforms to allow users to look through the socio-technical opacity of propagating information — in often limited attention span of user interaction on these platforms — and assess the credibility of that information?* into three more specific objectives that I address in Chapter 4, Chapter 5, and Chapter 6 in this dissertation:

- **RQ1:** As information surfaces within a platform (Google Search), how do different contextual factors behind that information impact its credibility?

- **RQ2:** When assessing credibility, how can easy access to information provenance on a platform (Twitter) and its contextual factors help users understand the socio-technical context of information provenance?

- **RQ3:** When moving to platforms beyond Twitter (TikTok), how do users employ the platform cues to identify the socio-technical context of information on that platform?

3.1 Answering RQ1: Audit-based content analysis to examine problematic content online

Algorithms of platforms like Twitter and Reddit facilitate the amplification of misinformation by bringing more user attention to it Fernández et al. (2021); Shepherd (2020). Researchers have employed auditing mechanisms to investigate the role of algorithms. Audits have shown how YouTube deploys algorithms with the potential to lure people down conspiracy *rabbit holes* by continuously suggesting related content (Rodriguez, 2018; Albright, 2018; Hussein et al., 2020). Auditing techniques have found that even e-commerce platforms like Amazon can promote a filter bubble effect, where users who browsed anti-vaccination content on the platform received relatively more suggestions promoting similar content than those who did not (Juneja and Mitra, 2021).

Researchers have illustrated the efficacy of the auditing technique in revealing intriguing and sometimes unintended platform behaviors. For instance, Epstein and Robertson (2015) demonstrated the existence of the search engine manipulation effect (SEME), wherein search engine providers can influence user behavior by manipulating search results. This manipulation was found to significantly sway voting preferences in favor of a particular candidate by as much as 20% or more (Epstein and Robertson, 2015; Spenkuch and Toniatti, 2016). Previous investigations into the behavior of the Google search engine have revealed that searching for specific queries with limited authoritative information, known as data voids, can result in the easy discoverability of conspiratorial websites (Bradshaw, 2019). Employing various interpretive techniques to analyze these patterns can provide deeper insights into such behaviors (Simko et al., 2021). For instance, researchers have utilized crowdsourced mechanisms to analyze data collected through search engine queries, shedding light on how Google's search technologies influence algorithmic information curation, particularly during the 2016 US elections (Diakopoulos et al., 2018).

To deepen our comprehension of how search engines may direct users to misleading information sources, I employed a similar approach by conducting a content analysis-based audit of Google's Search Engine Results Page (SERP) data, focusing on election-related content during the 2020 electoral period. To construct a comprehensive dataset concerning the political headlines during the 2020 US election season, I systematically gathered search headlines multiple times daily from October 5, 2020, to December 3, 2020. This effort

resulted in the scraping of over 800,000 headlines in response to 20 election-related keywords—10 general (e.g., 'Ballots') and 10 conspiratorial (e.g., 'Voter fraud')—across 20 cities spanning 18 states.

For the analysis phase, I utilized journalistic principles to develop a coding scheme, consulting with journalists and researchers to understand the headline construction process and the best practices related to online content dissemination and presentation. Drawing from this preliminary exploration, I structured the coding scheme around a central "Stance" category, facilitating the categorization of headlines based on their potential impact on users' trust in the election's legitimacy. Subsequently, I qualitatively coded a stratified sample of 5,600 headlines, focusing on the prevalence of misinformation (referred to as "delegitimizing information" within Chapter 4 for consistency with published work (Zade et al., 2022b)). I conducted several regression analyses to bolster the insights gleaned from the coding process. These analyses served to reinforce the observations emerging from the coding process, enhancing the robustness of the findings.

3.2 Answering RQ2: Research through design to investigate design conceptualizations in misinformation context

I adopted a research through design (RtD) approach to address the second research question (Gaver, 2012; Zimmerman and Forlizzi, 2014). RtD is a research method that leverages design practices to discover new knowledge (Zimmerman and Forlizzi, 2014). Researchers have shown how this approach can uncover insights into how variations in design impact users' affective and decision-making responses across different contexts, including the realm of misinformation (Carroll et al., 2020; Carroll and Bonkel, 2021; Sherman et al., 2021). In line with this approach, I aimed to curate a set of cues that offer rich context about the spread of a tweet, prompting participants to explore *how and why* they could use such interventions to assess the credibility of content when retweeting.

To help users understand the role of and context of different actors who are involved in spreading the information, I first came up with a framework that later guided the selection and design of the contextual cues. Borrowing from existing research and our own experiences, I conceptualized multiple possible cues — without being 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' (Boyd et al., 2010); using hashtags that consist of hate can signal the production of radicalized content (Agarwal and Sureka, 2015).

Next, I chose the different actors to be shown in the information trajectories that explain how information reached a user while balancing the amount of information to be shown to a user. I chose my first actor to be the known source of the tweet (root tweeter) given the established significance of primary source for online credibility assessment (Metzger et al., 2010; Geeng et al., 2020). Realizing the potential of top accounts with many followers in spreading content (Institute, 2021; Chong and Kim, 2020), I chose the 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 (Turcotte et al., 2015; Geeng et al., 2020), I chose an individual user's online friend as our third actor along the trajectory.

I then designed the novel intervention, which is highly relevant to providing socio-technical context and extensible for other researchers to build upon. To learn more about the effectiveness (or not) of the intervention towards discerning misinformation, I conducted an interview-based qualitative analysis of employing the intervention in everyday retweeting practices of regular users.

While Twitter serves as an excellent platform to test how several factors might impact a civic discourse that Twitter facilitates, these research efforts have often been criticized for not being representative enough with a bias towards political elites and individuals who tend to be politically more vocal (Morstatter et al., 2014; Bode and Dalrymple, 2016; Barberá and Rivero, 2015). It is important to note the presence of such a bias in the context of this research as it might impact the findings from the study and limit its generalization beyond the observed sample to some extent. To compensate for such a bias, I ensured the inclusion of a wide range of participants with diverse demographics in terms of education, political leaning, and media preference.

Next, I conducted an interpretive, grounded analysis of the data that was collected through participant interviews in the RtD exercise (Charmaz, 2006). Accordingly, I iteratively refined the themes that emerged from the analysis and recorded the insights using analytical memos. Finally, I reported on how contextual cues support participants in developing and/or refining their mental model about an online account and how the tweet trajectory facilitates making a quick but informed credibility judgment about the provenance of

information visible in the intervention. Given that a design intervention conveys a specific framing of the problem that we wish to explore, I note that knowledge generated through RtD is reflective of the functions and limitations of our intervention (as understood by the participants) (Zimmerman et al., 2007) and as imposed by the researchers (Dow et al., 2013).

3.3 Answering RQ3: Participatory design to generalize learning objectives for discerning problematic content across platforms

While numerous researchers have investigated platforms like Google, Twitter, and Reddit as socio-technical platforms, there is relatively little research available on information dynamics on TikTok due to its relatively recent emergence. However, the platform's rapid growth in users and user engagement makes it a compelling site for research.

For instance, Baumel et al. conducted a study comparing contrasting narratives, represented by hashtags WearaMask and MasksDontWork, in the most viewed videos on TikTok. They discovered a notable presence of medical professionals supporting masks as an effective strategy against Covid-19 (Baumel et al., 2021). In another study focusing on urinary tract-related content in 2022, TikTok was found to have significantly higher user engagement but less scientific information, lower credibility, and more misinformation compared to YouTube (Tam et al., 2022). Similarly, research has shown that anti-vaccination messages tend to go viral on TikTok, particularly among younger users, contributing to vaccine hesitancy (Basch et al., 2021). These findings underscore the urgent need for extensive efforts to counter misinformation spread on TikTok. In response to this challenge, several fact-checkers have joined the TikTok platform to debunk misinformation and combat its dissemination (Bautista et al., 2021).

In the previous study, I identified effective learning objectives that capture the socio-technical characteristics of information on Twitter. However, these lessons may not seamlessly apply to the TikTok platform due to its relatively unique and younger user base, often referred to as Generation Z (Rapkin, 2017). Therefore, I employed the participatory design method to adapt these learning objectives from Twitter to TikTok.

The participatory design approach allows participants to express their own identities, draw from personal experiences, and offer valuable insights to inform design mechanisms, particularly in the context of news

literacy initiatives (Literat et al., 2020, 2021). I utilized this approach to collaborate with participants to understand how they interpret and apply various problematic behaviors, similar to those identified in the context of Twitter, specifically to the TikTok platform. By identifying platform-specific signals that align with general learning objectives, we can facilitate incidental learning about the socio-technical complexities of information on online media platforms.

3.4 Reflecting on the overall methodological approach

Before delving into the studies conducted as part of this dissertation, it's important to consider the rationale behind selecting three different platforms and the unique methodological approach adopted for studying each platform.

Initially, the primary objective of this dissertation was to employ a mixed methods approach on the Twitter platform. The aims were twofold: firstly, to conduct a qualitative study by implementing an intervention that prompted users to consider the concept of provenance, and secondly, to conduct a quantitative experiment to measure the impact of the provenance intervention, utilizing a browser-based plugin, on how users adjust their credibility assessment and information-sharing habits. The findings from the latter experiment would complement those from the former. However, due to challenges encountered by Twitter, particularly with changes in leadership around 2022 and the subsequent loss of some users, I had to pivot and explore other platforms. While this decision affected the certainty of claims that could be made about a single platform, it allowed for a more comprehensive analysis and provided findings with higher generalizability.

Using three different methods—auditing, research through design, and participatory research—on three unique platforms had its own set of advantages and disadvantages, particularly regarding the validity of findings. In this dissertation, I ensured that I struck a balance between these through careful selection, implementation, and integration of diverse methods to study the potential utility of information provenance when making credibility assessments.

For example, while the auditing approach employed to study the Google Search platform lacked a user perspective, findings from research through design may have been influenced by the choice of design and cues used in the provocation. Conversely, findings from participatory research were influenced by the knowledge and experience of the participants. Collectively, synthesizing findings from these multiple approaches

was challenging as the studies engaged with the notion of provenance in very different ways. However, employing these diverse methods enabled me to delve deeper into the intricacies of the conceptualization of provenance. Carefully selecting the methods catered to the corresponding goals facilitated triangulating insights gathered through unique approaches to discovering knowledge. For example, in studies on the Twitter and TikTok platforms, findings from unique methodological approaches resonated with each other, thereby enhancing the validity of the judgment affordances framework presented in this dissertation. Thus, utilizing multiple methods together contributed to increasing the validity of findings.

Chapter 4

Witnessing information opacity on Google Search: Auditing headlines to examine search engine as a gateway to misleading content

This chapter presents the research published in the Journal of Trust and Safety in 2022 (Zade et al., 2022b), which I co-led with Morgan Wack, a co-author of this work. When using the pronoun 'we' within this chapter, it refers to all the co-authors involved in the broader project. I have adapted significant portions of this chapter from the publication, making some minor adjustments to the introduction. Additionally, I've expanded and refined the discussion section to provide contextualization relevant to this dissertation.

4.1 Introduction

Despite no evidence that widespread fraud occurred during the recent US elections (Cybersecurity & infrastructure security agency, 2021; Hale Spencer, Saranac, 2020; Cybersecurity & Infrastructure Security Agency, 2022), as reiterated in testimony by former Attorney General Bill Barr (Thompson et al., 2022), there remains skepticism among the public about the legitimacy of the election results. Following the elec-

tion, nearly 65% of Republican voters believed that the results of the 2020 US general election were illegitimate (Pennycook and Rand, 2021a). Such skepticism isn't unique to 2020 elections; during the 2018 midterm elections, voters who cast their votes using mail-in ballots were skeptical that their votes would be counted correctly (Alvarez et al., 2021). Though considerable effort has been spent studying how social media platforms serve to connect people to conspiracy theories, rumors, and misinformation related to unsubstantiated voter fraud, less is known about how and what kind of political content is spread through search engines.

Search engine platforms serve as the doors to information and news on the internet. In 2020, 65% of Americans used search engines as a primary source to gather news and information Shearer (2021), of which Google has a global market share of over 90% (StatCounter, 2021). As evidenced by "election results" and "coronavirus" constituting the top two search terms on Google in 2020 (Google Search, 2020), search engines have a tremendous potential to provide access to critical information that can influence democratic discourse. This is particularly true during election periods—in particular, the 2020 US general election—when political polarization, COVID-19 uncertainty, and demand for election information were all high (Kapferer, 1987; Bordia and DiFonzo, 2017; Starbird et al., 2020).

The 2020 US election gave rise to several narratives that cast doubt on the legitimacy of the results. Several official organizations, including the Cybersecurity and Infrastructure Security Agency (CISA), have debunked these narratives, and CISA confirmed in December 2020 that it was indeed a "secure election" Cybersecurity & infrastructure security agency (2021); Cybersecurity & Infrastructure Security Agency (2022); Hale Spencer, Saranac (2020). Despite the acknowledgment of confidence in the election by several government officials and elected leaders, both Democratic and Republican (Brennan Center for Justice, 2020), unproven and misleading election-related narratives were (and some remain) widely available online. Users accessed these narratives by means of different information provenances, *e.g.*, provenances that originated from legacy news channels like CNN and Fox News, provenances that consisted of prominent Twitter accounts, etc. Soon, search engines might also give access to TikTok accounts that popularize the information, highlighting the role of these TikTok influencers in the information provenances. To what extent do these provenances vary in terms of how credible information they serve? In addition, the credibility of provenance, as perceived by users, might also be impacted due to other characteristics like search location, time of the

search, etc., as they provide different contextual insights into that provenance. In this Chapter, I examine (*RQ1*) of this dissertation: as information surfaces within a platform (Google Search), how do different contextual factors behind that information impact its credibility? More specifically, I report my investigation on *whether and potentially how Google served as a gateway to content that may have undermined trust in election processes, institutions, and results*. We audited headlines appearing in Google's SERPs in response to several search terms before, during, and after the 2020 US election, focusing on provenance. Specifically, our research was guided by the following smaller questions, each of which addresses a unique contextual characteristic of provenance:

- *Question One:* How do the SERP verticals—search results, stories, videos, and advertisements—differ in the amount of misleading content?
- *Question Two:* How does one's location in a specific city—split by population and party representation—change the kind of election content found in search results?
- *Question Three:* Do different search terms lead to different search result quality?
- *Question Four:* Which online news domains served as the most frequent gateways to content that may have undermined trust during the election period?

To answer these questions, we focused on news headlines from Google's SERP data (see Figure 4.1). The headline of a news story is known to influence users' interpretation of the story's content (Tannenbaum, 1953) and impact its popularity (Rieis et al., 2015). We collected headlines using election-related search keywords as seen on Google's search engine across 20 locations spread throughout the US. Since Google does not officially support a search API, and other services do not support location-specific requests, we resorted to a third-party paid service called SerpApi (SerpApi, 2020). This service allowed us to perform searches such that the results were associated with the locations of our 20 selected sites, rather than the results that Google would normally associate with the geographic location of our local IP address. Our data collection began before the election in early October 2020 and ran through mid-December. We performed an extensive qualitative analysis of a random sample of 5,600 headlines from over 500,000 SERP search results, 242,000 SERP stories, 62,000 SERP videos, and 47,000 SERP advertisements to evaluate the potential of

SERP data to undermine trust in the election. In addition to the analysis, we make the raw Google SERP data corresponding to election-related keywords across several disparate locations openly available to further analysis by other researchers (Zade et al., 2022a). [1]

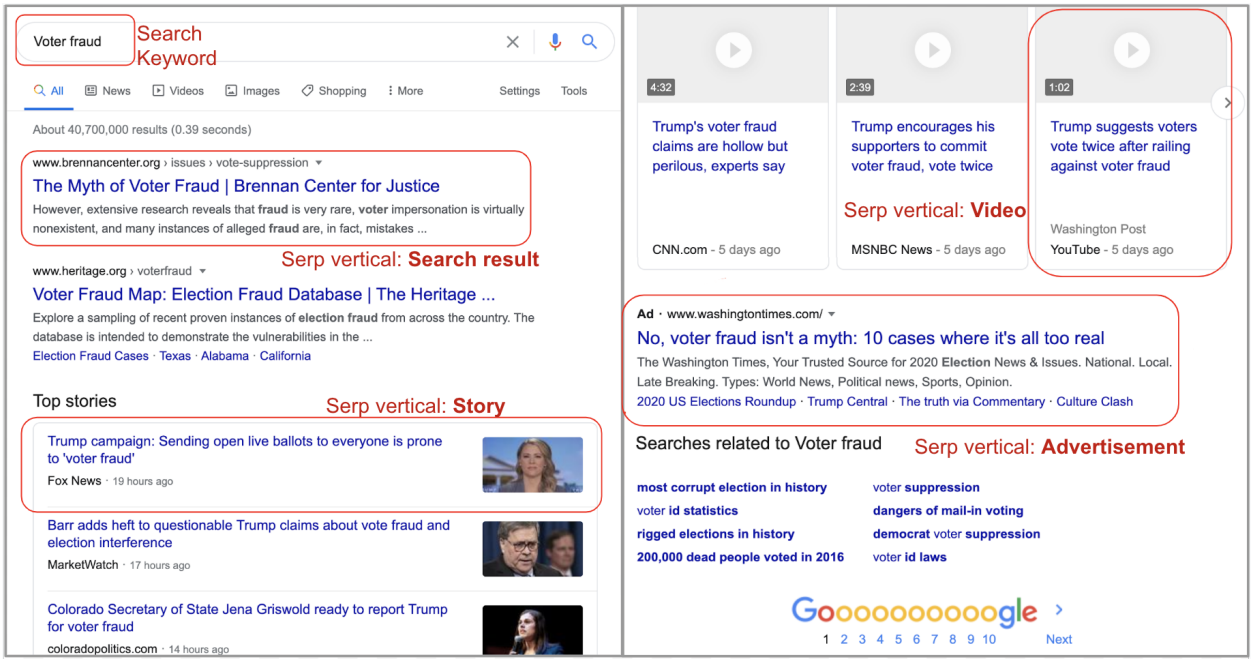


Figure 4.1: Two example screenshots of Google SERP data if a user were to search for "voter fraud." Left: We collected the headlines and metadata for the search results and all three top stories. Right: We also collected the headlines and metadata for all three videos and the only advertisement. Overall, we collected the first ten search results, top ten stories, top ten videos, and all the advertisements returned by the search engine in response to all keywords.

From these searches and our subsequent coding of reported headlines, we found that the headlines of the video content reported in our Google Search results pages contained a disproportionate amount of undermining-trust content when compared to alternative SERP verticals (search results, stories, and advertisements). Although swing states received more campaign advertisements than non-swing states, a user's location generally did not moderate the quality of information served by search engine headlines. We also found that the headlines displayed on the results pages were more likely to undermine trust when searches included conspiratorial election-related terms (e.g., "Voter fraud," "Rigged election," etc.) as opposed to general election-related terms (e.g., "Ballots," "Where do I vote," etc.), as well as if the headlines were as-

[1] Data is available on the Open Source Foundation platform (Zade et al., 2022a).

sociated with media domains with a relatively more right-leaning bias. Upon investigating the mainstream media headlines specifically, we found that legacy news sites with large audiences like The Washington Post and Fox News played an outsized role in delivering content with the potential to undermine trust. Finally, we present the topics that were the focus of trust-undermining and trust-imparting content across our coded sample.

Our study builds upon previous work that has emphasized the influence of online electoral content in altering perceptions about the legitimacy of the 2020 US general election. First, we present a novel dataset consisting of geographically and topically distinct search results presented by Google prior to, during, and following November 3, 2020. Second, we developed a coding scheme for assessing headline content in SERP data and its role in undermining trust that can serve as a template for future studies. Third, using our coding scheme, we analyze the political content likely presented to a large number of users on Google's platform before, during, and shortly after the election. From this analysis, we identify the topics, domains, and search patterns of election-delegitimizing content. In particular, we present evidence that information with different provenances can vary in their credibility. We conclude with recommendations focused on open, auditable, and anonymized data for investigating these research questions in future elections.

4.2 Background: 2020 Election Delegitimization

4.2.1 Significance of news headlines

News headlines, along with other forms of content collected in Google Search results, play a critical role in conveying information and creating impressions. For news headlines specifically, a survey conducted nearly a century ago found that out of 375 people, 192 based their opinions about news from the reading and skimming of the headlines only (Emig, 1928). The importance of headlines in content conveyance has persisted as news has shifted online. Psychologists have known that early impressions matter and that early biases affect users in what they learn in further impressions of the artifact (Digirolamo and Hintzman, 1997). Based on analysis of about 70,000 headlines, Reis et al. confirmed that the sentiment of the headline could have a serious impact on how popular the story might become and the kind of discourse it encourages (Rieis et al., 2015). These projects have reiterated how headlines can serve as influential shortcuts for readers that

can subsequently guide their interpretation of the news (Tannenbaum, 1953).

Misleading content, even if only slightly misleading, can bias the interpretation of events, such as elections. This is why they are often used to frame real-world events in a particular light Jamieson et al. (2007); Liu et al. (2019). Framing strategies have often been employed—as was tracked during the 2004 Canadian federal election—to select aspects of particular news stories that increase the salience of the writer's or news source's chosen perspective (Andrew, 2007). By inducing bias in readers, exposure to misleading headlines can limit the capacity of its audience to process corrected information, thereby impacting their memory and reasoning (Ecker et al., 2014). Complicating matters further, readers have a tendency to over-weight headlines that are consistent with their social and political attitudes (Beam, 2014) while choosing to focus on headlines that they perceive to be true a priori (Edgerly et al., 2020), leaving readers vulnerable to misleading headlines that align with partisan values. The challenge posed by misleading headlines has been exacerbated by the growing use of social media platforms, where headlines are often prominently displayed as a substitute for the actual content of the article Gabielkov et al. (2016). In fact, there is little incentive for platforms to push users off the platform to the actual article. Despite these growing concerns, little is known about the role of headline content appearing in different SERP verticals (e.g., stories versus videos) during elections to undermine voter trust. This is the focus of our research.

4.2.2 Role of Google Search in shaping user opinion

Google Search is the most commonly used search engine StatCounter (2021) and, therefore, the focus of numerous studies into search engine function and performance. A recent study found that Google fares better in limiting the promotion of conspiratorial results as well as the presentation of links to conspiracy theory-dedicated websites when compared with other search engines like Bing, DuckDuckGo, Yahoo, and Yandex (Urman et al., 2022). Despite relatively higher resilience to conspiratorial content, concerns remain regarding bias evident in Google Search results Robertson et al. (2018). These potential biases are of concern to election integrity advocates, who have shown that Google's search engine has in the past privileged certain topics on its News homepage (including a disproportionate presentation of articles detailing the 2016 Trump campaign over his challengers) (Diakopoulos et al., 2018).[2]

[2]Diakopoulos et al. found that during the 2016 US elections, Google News had 941 indexed articles about Donald Trump, 710 about Hillary Clinton, and 630 about Bernie Sanders (Diakopoulos et al., 2018).

When investigating Google's role in shaping user attention to the news, Trielli and Diakopoulos (2019) found a small skew towards the political left in Google Search results. Although the diversity of the media sources varied by topic, a small fraction of the media contributed about 50% of the overall suggestions in the top stories. Similarly, recent research has found that a small number of sources contributed the majority of the stories about the 2020 US presidential election on Google SERPs (Kawakami et al., 2020). Epstein and Robertson (2015) showed that a search engine manipulation effect (SEME)—i.e., influencing user behavior through manipulation of search results by search engine providers—can impact the outcomes of elections. Voting preferences can be strongly influenced in favor of a candidate (20% or more) by showing search results biased toward a particular candidate (Epstein and Robertson, 2015; Spenkuch and Toniatti, 2016).

Search engines can impact user perception about the credibility of the news not only through the selection of stories (and sources) on the results page but also through the rankings in which these stories appear. A higher position in the ranking of a (SERP vertical) story impacts user decisions more, even if it is less relevant to the topic of the user's search, than another story that appears at a lower rank (Pan et al., 2007). Researchers have also questioned the role played in content presentation across different information modalities including text, stories, and videos across several platforms. Though recent work has suggested that video content may not be as persuasive as was once feared, users tend to believe in a video more easily than in text (Wittenberg et al., 2021). Given the increased prevalence of video-based misinformation, there is a shared belief among researchers that the real extent of persuasiveness of videos might diverge in real settings that are not lab-controlled. For example, when comparing the role of text versus video modality within messaging apps, researchers found that users process videos superficially and tend to more influenced by them compared to text (Sundar et al., 2021). Based on this result, we compare the different SERP verticals—e.g., news, stories, search results, and ads—in our study.

4.2.3 Auditing as a method to trace mis- and disinformation

Algorithms of platforms like Twitter and Reddit facilitate amplification of problematic content by bringing more user attention to it Fernández et al. (2021); Shepherd (2020). Researchers have employed auditing mechanisms to investigate the role of algorithms. Audits have shown how YouTube deploys algorithms with the potential to lure people down conspiracy "rabbit holes" by continuously suggesting related content (Ro-

driguez, 2018; Albright, 2018; Hussein et al., 2020). Auditing techniques have found that even e-commerce platforms like Amazon can promote a filter bubble effect, where users who browsed anti-vaccination content on the the platform received relatively more suggestions promoting similar content than those who did not (Juneja and Mitra, 2021).

Researchers have expressed hope that the auditing method can enable us to witness and understand why some unwanted platform behaviors occur Simko et al. (2021). For example, prior investigation focused on understanding Google search engine's behavior has shown that searching for specific queries that have limited authoritative information (i.e., data voids) can lead to easy discoverability of conspiratorial websites Bradshaw (2019). In order to expand our understanding of how search engines can lead users to misleading content, we conduct an audit of Google SERP data focused on election content during the 2020 electoral period.

4.3 Data Collection

4.3.1 Search terms

To conduct our analysis, we generated a list of election-related search terms in October 2020 (see Table 4.1). These terms were used to assess differences in headlines related to different SERP verticals: search results, news, advertisements, and videos. We split our terms into two distinct categories. The first category of search terms aimed to capture the results produced when searching for general election-related content. This included terms such as "presidential election" as well as common election questions such as "where do I vote." We also included a second category of terms targeting electoral conspiracy theories identified across existing misinformation narratives. This list was designed to mimic potential searches focused on issues related to the legitimacy of election processes and results. As the list was developed in advance of the election in September, it was informed by prior political controversies and online rumors and does not include terms related to conspiracy theories such as Sharpiegate, which only became relevant after election day.[3] As such, it was comprised of both general conspiratorial phrases such as "election fraud" and "stolen

[3]The 2020 Sharpiegate conspiracy theory, which claimed that sharpies were deliberately distributed to Republican voters in order to invalidate their votes, is distinct from the prior controversy related to Donald Trump's use of a sharpie on a weather map displaying the trajectory of Hurricane Dorian in 2019.

election" as well as more specific actions such as "voter fraud" and "ballot dumping."

General Terms	Conspiratorial Terms
Election results	Rigged election
Ballots	Late ballots
How do I vote	Voter fraud
Where do I vote	Voter intimidation
Mail-in voting	Election fraud
My ballot	Electoral fraud
Absentee ballot	Stolen election
Presidential election	Ballot harvesting
Vote by post	Ballot dumping
Vote	Mail dumping

Table 4.1: Election-related search terms fed into Google's search engine. Ten of the search terms were general election terms, and the other ten terms were linked to conspiracy theories related to the 2020 US presidential election.

4.3.2 Search locations

Google customizes its search results based on geographic location Rogers (2013). The results of a search for the terms "election results" in Los Angeles, California, for example, could be different than the results of the same search in Topeka, Kansas. These differences can, in turn, shape geographic differences in how individuals think and behave, since search results can both prime audiences to think about certain issues and frame how they think about those issues Zook and Graham (2007). However, the exact relationship between search customization and local understandings of emerging news events remains understudied Ballatore et al. (2017). To contribute in this area, we developed a purposive sampling approach to collect search results across locations in the US that varied by region and degree of urbanization. Social scientists have long explored how shared economies and cultural traditions produce regional sociopolitical identities, and urban-rural divides have emerged as an even more salient variable in shaping current partisan politics in the US Gimpel et al. (2020).

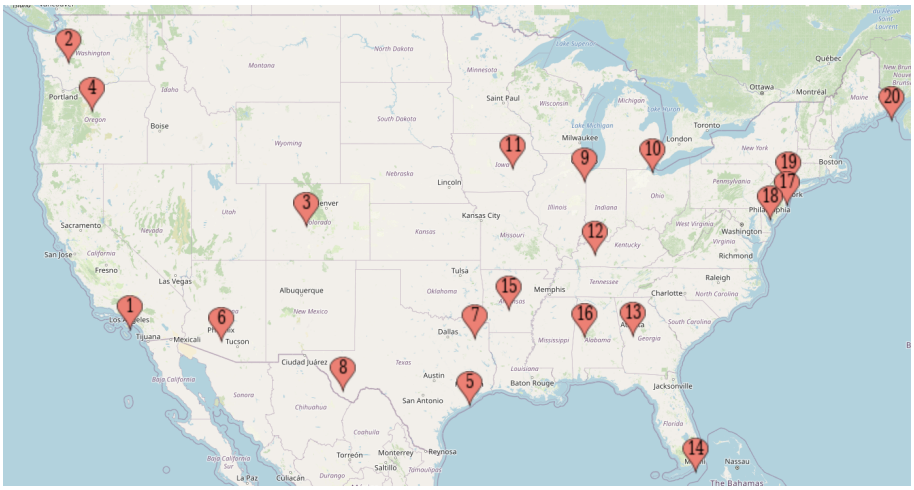


Figure 4.2: Geographic spread of the 20 locations across which we scraped Google Search results for search terms listed in Table 4.1.

MapID	City, State	Size	Swing	MapID	City, State	Size	Swing
1	Los Angeles, CA	UA	N	11	Cedar Falls, IA	UC	N
2	Seattle, WA	UA	N	12	Santa Claus, IN	RA	N
3	Vail, CO	UC	N	13	Atlanta, GA	UA	Y
4	Grass Valley, OR	RA	N	14	Miami, FL	UA	Y
5	Houston, TX	UA	N	15	Morrilton, AR	UC	N
6	Phoenix, AZ	UA	Y	16	Berry, AL	RA	N
7	Clarksville, TX	UC	N	17	New York, NY	UA	N
8	Fort Davis, TX	RA	N	18	Philadelphia, PA	UA	Y
9	Chicago, IL	UA	N	19	Poughkeepsie, NY	UC	N
10	Detroit, MI	UA	Y	20	Eastport, ME	RA	Y

Table 4.2: Our data collection includes Google SERP data as rendered in these 20 cities spread across 16 states in the USA. *UA* refers to urban areas, *UC* refers to urban clusters, and *RA* refers to rural areas. *Y* or *N* refers to whether it was a swing state or not.

To select locations, we first divided the US into Northeast, Southeast, Southwest, Midwest, and West regions, drawing on a common five-region classification schema (National Geographic, 2009). Within each of those regions, the project identified four locations that represented varying levels of urbanization. Here we used the U.S. Census Bureau's classification of locations as urbanized areas (UAs) of 50,000 or more people; urban clusters (UCs) with populations between 50,000 and 2,500; and rural locations (designated as "rural areas", or RAs, in this study) that have a population under 2,500 US Census Bureau (2010). For

each region, we chose two UAs, one UC, and one RA. We chose to overrepresent UAs because there tends to be more election-related news and activity in more densely populated locations, allowing us to better examine regional differences across these larger markets. However, we also attempted to select UAs within each region such that they also varied in size, with one containing a population of several million and the other a population close to one million. We selected specific locations within this framework, based on our knowledge of interesting news having emerged from those locations, as shown in Figure 4.2. Our hope was that this would produce a richer dataset. We also strove to select locations that were diverse in partisan political orientation. In many instances, our first-choice rural areas were not found within the API we used for data collection. In these instances, we chose a nearby city that could be found within the API. This process resulted in 20 locations, as listed in Table 4.2.

4.3.3 Search service

Google does not officially support any search API, and other search services do not allow easy access to location-specific SERP data. While we had access to a white-listed IP address to crawl unlimited Google SERP data, this data would have reflected SERP results as seen from that specific location. To accommodate location as a factor in SERP-related audits, prior research resorted either to using browser-based plugins (Robertson et al., 2018) (limiting the data collection to queries adopted by select users at specific times), or to making data requests from multiple locations with unique IP-addresses (Mustafaraj et al., 2020) (limiting the scalability to only a few unique locations). To overcome these limitations, we used the SerpApi platform (SerpApi, 2020) to search for keywords of our choice mentioned in Table 4.1 at regular intervals each day and fetched the corresponding Google Search results as it would be seen at the 20 unique locations listed in Table 4.2.

The SerpApi API provides real-time scraping of Google SERP data, allowing users to select a specific search location from available choices without adjusting based on the researcher's IP address. While the location-specific search feature is unique to SerpApi, we acknowledge that any biases in this data could affect research findings. To validate the data collected using SerpApi, we conducted two checks. Firstly, we compared SerpApi data with data displayed on Google for generic keywords such as "School," "Cafe," "Museum," etc. We found that the SerpApi data matched exactly what was seen on Google Search, but

this confirmation was limited to the Seattle-specific location, as Google Search doesn't offer the option to retrieve data from different locations. Secondly, we examined location-specific differences in SERP data returned by the API. Using the same generic keywords, we confirmed variations depending on the specified location, even for election-related searches. For instance, searching for "Vote" yielded different headlines: "How to vote the new way in L.A. (in 2020)" for Los Angeles, California, and "How to Vote In Colorado" for Vail, Colorado.

Upon confirming the validity of the SerpApi data, we utilized the paid version of the API to conduct approximately 50,000 unique searches per day. This data is openly accessible on the Open Source Foundation platform for future research projects (Zade et al., 2022a).

4.3.4 Search schedule

We intended to scrape the SERP data several times a day to capture news headlines soon after they were released by different media sources. As we began the collection, we collected data four times every day (3:00, 9:00, 15:00, 21:00 EST) between October 5 and October 29, 2020. Later, we reduced this frequency to three times a day (00:00, 08:00, 16:00 EST) from October 30 to December 3, 2020, to fit within the constraint of 50,000 allowed searches based on our service subscription and required searches for a related project. Even with the reduced frequency, we were able to capture news headlines in the morning, evening, and late night (EST) as intended.

4.3.5 Overall collection

For every search, we collected the first ten search results, top ten news stories, top ten videos, and all advertisements returned by the search engine in response to a search keyword, which is more than the information rendered on the Google SERPs as seen by the user and illustrated in Figure 4.1. It included the headlines of all the components and corresponding attributes, such as website link, domains, date and time of publishing (for videos and stories), etc., as seen on the Google search engine. Overall, our initial collection consisted of 56,763 unique location-specific keyword searches. Across these searches, we collected about 47,000 advertisements, 500,000 search results, 240k,000 stories and 66,000 videos.

Given that higher-ranked results are known to influence user decisions Joachims et al. (2007); Brooks

(2004); Lorigo et al. (2008), we decided to focus on the top five search results, top three news stories, top three videos, and all included advertisements. Focusing on the higher-ranked results across varying SERP verticals—comprising 485,805 results—allowed us to inspect headlines that had greater influence on user opinions. These contained 47,000 advertisements (same as before since we always considered all the advertisements), 283,000 search results, 242,000 stories, and 36,000 videos.

For each combination of search keyword and search location, we now had either three or four SERPs per day depending on the frequency of collection during that time. For each of those combinations, we then randomly selected one SERP per day to make the data sample size more manageable and ensure even distribution of headlines across the duration of two months. This reduced our sample to 174,511 total headlines including about 14,000 advertisements, 97,000 search results, 41,000 stories, and 20,000 videos. Table 4.3 summarizes the steps we took to filter the sample of headlines.

Step#	Procedure	Resultant data sample
Step 1	We collected SERPs (10 search results, 10 stories, 10 videos, and all ads) for 20 search keywords (Table 4.1) as seen at 20 locations (Table 4.2) several times a day using SerpApi.	56,763 unique location specific SERPs; about 47k ads, 500k search results, 240k stories, and 66k videos.
Step 2	To focus on data that easily appears on SERPs without any extra user clicks, we selected the top 5 search results, top 3 stories, top 3 videos, and all ads.	About 47k ads, 283k search results, 242K stories, and 36k videos.
Step 3	For each combination of search keyword and search location, we randomly chose exactly one SERP per day.	About 14k ads, 97k search results, 41k stories, and 20k videos. Summary statistics in Table 4.4.
Step 4	Using stratified random sampling technique, we split the Oct.–Dec. 2020 duration into four 2-week long periods and selected 50 SERP headlines per combination of location type (2 urban areas, 1 urban cluster, 1 rural area), SERP vertical type (result, stories, videos, ads) and search term type (general, conspiratorial).	1,600 stories, 1,600 videos and 1,600 searches across 4 time periods and 800 ads across the first 2 time periods; out of the 5,600 SERP headlines (as per power analysis), we qualitatively coded 2,438 unique ones.

Table 4.3: Step-by-step description of how we sampled the headlines in our SERP data to make it suitable for qualitative coding.

4.3.6 Filtered collection for qualitative coding

: After we collected the data, we assigned a label and coded each headline into different categories. Although the same headline could appear multiple times in our data—e.g., the headline "Voter Fraud Map: Election Fraud Database" appeared once in relation to Atlanta, Georgia, and then in relation to Cedar Falls, Iowa—we only coded unique headlines. To filter the 174,511 headlines and generate a set small enough for manual coding but large enough to allow the use of inferential statistics, we conducted a power analysis using the G-power tool (Faul et al., 2007). Given that the assigned codes served as the outcome variables, we chose a two-tailed a priori analysis for the z-test family suitable for logistic regression and discovered that we needed a sample size of 5,408 headlines—assuming a minimal effect size corresponding to an odds ratio of 1.1 with about 80% power.

Median number of headlines per day across 20 locations.				
Search Keyword	*Search Results* (Top 5)	*Stories* (Top 3)	*Videos* (Top 3)	*Advertisements* (All)
Absentee ballot	100	54	0	55.5
Ballot dumping	100	0	3	1
Ballot harvesting	100	57.5	13.15	3.5
Ballots	100	60	45	38.5
Election fraud	100	60	6	44
Election results	100	60	28.5	22
Electoral fraud	100	60	0	25
How do I vote	100	0	0	32.5
Late ballots	100	56.5	0	14
Mail dumping	100	0	57	0
Mail-in voting	100	60	25.5	51
My ballot	100	28.5	0	14
Presidential election	100	60	51	19.5
Rigged election	100	60	15	55.5
Stolen election	100	55	1.5	20
Vote	100	60	30	25.5
Vote by post	100	0	0	25.5
Voter fraud	100	60	24	57
Voter intimidation	100	54	9	3
Where do I vote	100	0	0	27.5

Table 4.4: Summary statistics of SERP data separated by SERP verticals and search keywords. Given the skewed nature of the data—e.g., while searching for "Electoral fraud" returned a maximum of 75 ads (October 5, 2020), searching for "Ballot dumping" only returned a maximum of three ads (October 8, 2020) across different locations—we report the median measure as our choice of summary statistic. A median score of 0 indicates a relatively lesser (but non-zero) number of headlines for the corresponding keyword.

To ensure that the data evenly represented the different search terms, search locations, and information modalities, but was not biased either by the time or day when it was scraped, we opted for a stratified random sample. We split our timeline into four 2-week long periods—Oct. 5–19, Oct. 20–Nov. 3, Nov. 4–18, and Nov. 19–Dec. 3—such that each period contributed evenly to our sample. We next set out to select 50 search instances per combination of city type (two urban areas, one urban cluster, one rural area), SERP vertical type (result, stories, videos, ads), and search term type (general, conspiratorial)—thus, selecting 1,600 headlines for each of the four time periods that will overall exceed the sample size of 5,408 as suggested by the power analysis. Unfortunately, we could not fetch 1,600 headlines from advertisements since (1) there were no advertisements for any of the issue-specific terms in the third and fourth time period

after November 4, and (2) Google did not surface 50 advertisements per city type even for the regular search terms. To overcome this asymmetry, we collected ads for only the first and second time periods. Our data sample thus consisted of 1,600 stories, 1,600 videos, and 1,600 searches across four time periods and 800 ads across the first two time periods.

These 5,600 headlines selected through a stratified random sampling were not necessarily unique. For example, the headlines "Mail carrier arrested for dumping mail" and "USPS employee arrested, accused of dumping mail" showed up the most—61 and 59 times, respectively—at different locations and/or in different search batches within our sampled set. We then coded the unique 2,438 headlines out of this set using the codebook described below.

4.4 Coding Scheme

In designing the coding scheme, initial data was first analyzed during a two-week exploratory period. During this time, we spoke with journalists and researchers about the headline construction process, including discussion of best journalistic practices related to the dissemination and presentation of online content. These practices included an emphasis on centering facticity through the use of keywords associated with falsehoods (e.g., "misinformation," "false accusations," "misleading"), avoiding the spotlighting of problematic groups, focusing headlines on impact rather than eventizing aberrations or anecdotes, and ensuring that headlines are well-matched with the content of the related article rather then solely matching on prominent terms. In addition to providing insights such as these to help inform the coding scheme, the preliminary period also allowed us to simplify the primary categories in our coding scheme.

Informed by this preliminary process, we developed the coding scheme around a central "Stance" category, which was used to categorize headlines based on their potential impact on search engine users' trust in the election's legitimacy. Once this central variable was in place, we trained three coders to differentiate between various codes on this dimension, which sought broadly to answer the question:

If voters were to have read this headline on the day it was captured, how (if at all) could it have affected their perception of the integrity of the 2020 US election's processes, institutions, and results?

Eventually the "Stance" category was narrowed to focus on three central codes: *Sows Doubt*, *Imparts Trust*, and *Provides Information*. The shortened definitions of these separate codes were finalized as follows:

- **Sows Doubt:** The headline has the potential to lower voter trust in the election's integrity.

  Example: "Allegheny County ballot contractor accused of sending out late ballots in other counties"

- **Imparts Trust:** The headline has the potential to improve voter trust in the election's integrity.

  Example: "Barr says he hasn't seen fraud that could affect the election outcome"

- **Provides Information:** The headline is not likely to alter voter trust in the election's integrity.

  Example: "Biden projected to win Georgia, Trump projected to win North Carolina."

In addition to these three central codes, we added two more codes to the "Stance" category. The *Campaign Ad* code was used to identify content that appears in SERP verticals like search results or videos but were merely a promotional campaign in nature. The *Other* code identified headlines that did not pertain to the election at all.

Based on insight from the initial exploratory period illustrated that headlines coded as either *Imparts Trust* or *Sows Doubt* could be further divided to discern headlines actively spotlighting or emphasizing issues related to the election's legitimacy. For example, while one subset of headlines was coded solely as *Sows doubt* (or *Imparts trust*), denoting its potential to reduce (or impart) trust in election integrity—e.g., "Poll worker accused...," "voters are concerned..."—a second subset appeared constructed specifically to undermine (or bolster) perceptions of its integrity—e.g., "Voter Fraud Map: Where to find evidence...," "6M+ votes shifted by big tech... ." To capture this crucial difference, the "Promotion" category (which involved a binary code) was developed to augment the "Stance" categorization. Collectively, the two categories are reported in tandem to ensure that we identify not only headlines that promote distrust, but also those that promote trust in the election's legitimacy.

In cases where we assigned the "Stance" as *Imparts Trust*, the "Promotion" category was used to identify headlines that deliberately attempted to build trust in the integrity of the election among readers. Similarly, where the "Stance" was coded as *Sows Doubt*, the "Promotion" category was used to identify headlines that appeared to be deliberately aimed at undermining perceptions of the election's integrity. Definitions and examples of headlines that we determined to promote distrusting content are presented below:

- **"Promotion" + "Sows Doubt" = "Promotes Distrust"):** The headline is *actively* reducing voter trust in the election's integrity

Example: This accounts for differences in headlines discussing topics that may undermine trust in the election, such as "Voters fear voter suppression in the build-up to the election," and headlines that push these narratives, such as "Guns, lies and ballots set on fire: This is voter suppression in 2020."

- **"Promotion" + "Imparts Trust" = "Promotes Trust"):** The headline is *actively* improving voter trust in the election's integrity

  Example: This accounts for differences in headlines discussing topics that may improve trust in the election, such as "Ohio county officials shoot down Trump claim of 'rigged election," and headlines that push narratives to improve trust, such as "Election fraud claims are baseless."

Content coded as both *Promotion* and *Sows Doubt*) is the closest to matching our conception of content with the potential to undermine trust in the election. As such, we used this subset as the basis for the primary analyses included here. Additional categories were included in the coding scheme, but remain peripheral to the central analyses discussed in this paper. These are discussed further in Appendix **??**.

4.4.1 Coding Process

Once the coding categories were finalized, a subset of 200 of the 5,600 selected headlines were used as a practice set to test out the final coding scheme on real data. Once each coder had completed their coding of the initial set, the lead researcher on the project went through each disagreement individually with all three coders to identify issues in the coding scheme to ensure consistency across the coders before moving on to the full set. Most of the discrepancies resulted from differences in each coder's knowledge of the conspiracy theories that had proliferated online during the election period, which resulted in more knowledgeable coders correctly identifying headlines coding these narratives as "Promotes Distrust." Less knowledgeable coders were subsequently given a longer list of common conspiratorial narratives to review.

Once the coding scheme was finalized and the coders felt confident in their ability to discern between the codes in each of the categories, the data was organized in descending order based on the frequency of headline appearance in the database. This resulted in the collection of 492 headlines that occurred more than two times each in the database. All three coders coded them as a final check to ensure shared understanding of the coding scheme. After arbitrating any coding conflicts and determining enough consistency across coding, the team then proceeded to code the entire primary headline dataset.

The unique 2,438 headlines were randomized and each coder was given 2/3 of the headlines to code, resulting in each headline being coded twice by two different coders. After the first two coders finalized their coding, we found that our coders shared an *almost perfect* understanding of the "Stance" and "Promotion" categories as indicated by a Cohen's kappa of 0.78 and 0.9 respectively (Landis and Koch, 1977). Any disagreements between the first two coders were then arbitrated by a neutral third coder.[4]

4.5 Results

4.5.1 R1: SERP vertical type

Our analyses show strong correlations between specific SERP verticals and the frequency of headlines that promoted distrust in the election's integrity ("Promotes Distrust"). Specifically, as seen in Figure 4.3, videos during the period were more likely to contain undermining content than other SERP verticals by a wide margin. This relationship persists both with headlines that serve to sow doubt in the credibility of the election and also among the more concerning content that promotes, rather than simply discusses or mentions, similar content.

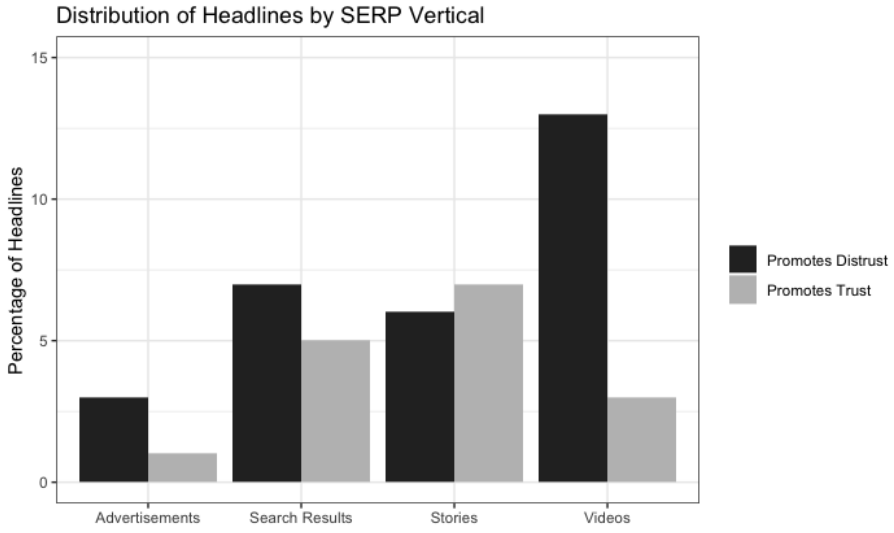


Figure 4.3: Percentage of coded headlines that promoted trust and distrust in the integrity of the election.

[4]Overall, the final agreement rates ranged from 75% to 99%—corresponding to a Cohen's kappa of 0.69 and 0.99—suggesting shared understanding across all the coding categories. We have included the inter-coder reliability measures across all the categories in Appendix **??**.

We ran a multinomial regression analysis by modeling the SERP vertical type (advertisement, search results, top stories, videos) to compare the extent to which headlines promoting distrust and promoting trust were identified in headlines for videos and top stories. We found that the odds of a video having a headline containing content with the potential to undermine trust were almost three times greater than headlines associated with a story (see Table 4.5). Moreover, top stories were about three times as likely to promote a trust-imparting headline than videos, suggesting that video headlines contained both disproportionately high amounts of content that promoted distrust as well as low amounts of content that promoted trust.

SERP vertical type	Odds ratio	CI [95%]	p-value
Sowing doubt and promoting it (Yes, No; reference: No)			
(Intercept)	0.049^*	[0.036, 0.069]	$< .001$
Searches	2.213^*	[1.536, 3.186]	$< .001$
Stories	1.874^*	[1.294, 2.713]	$< .001$
Videos	5.472^*	[3.867, 7.741]	$< .001$
Imparting trust and promoting it (Yes, No; reference: No)			
(Intercept)	0.009^*	[0.004, 0.019]	$< .001$
Searches	6.991^*	[3.227, 15.144]	$< .001$
Stories	8.755^*	[4.062, 18.869]	$< .001$
Videos	2.905^*	[1.295, 6.514]	$< .001$

Table 4.5: Odds ratios for "Sowing doubt and promoting it" and "Imparting trust and promoting it" through different information modalities of searches, stories, and videos (over campaign ads) calculated using logistic regression.

Figure 4.3 illustrates how top stories were the most common channel for the promotion of content that served to enhance readers' trust in the integrity of the election. When compared with other modalities such as videos, ads, and search results, stories were the only SERP verticals with more headlines that imparted trust than headlines that sowed doubt throughout the sample. When viewed longitudinally in Figure 4.4, we see that this discrepancy between SERP vertical type and trusted content was more prominent in the post-election period. We found an increase in the post-election videos with headlines like "Dominion whistleblower says she didn't see a single vote cast for..."[5] and "ELECTORAL FRAUD: Where To Find The Evidence | Rudy Giuliani | Ep. 89."[6]

[5]The entire title read as "Dominion whistleblower says she didn't see a single vote cast for Donald Trump in her 27 hour shift" and directed users to a YouTube video that can still be accessed online as of April 30, 2022.

[6]This video was later removed from YouTube for violating its community guidelines.

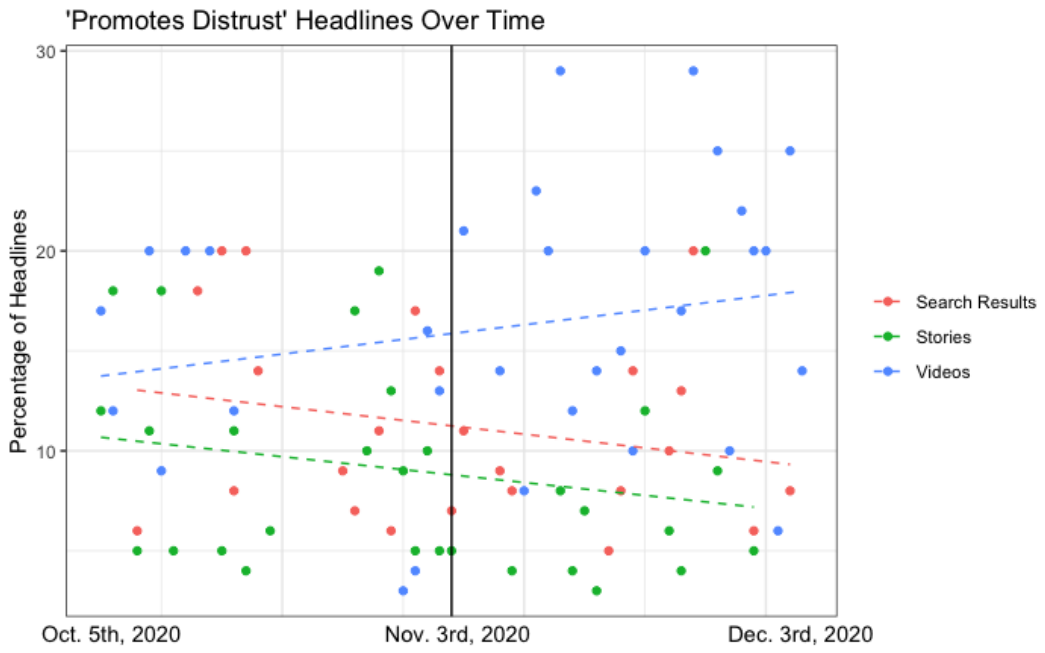


Figure 4.4: Percentage of headlines per day (Y-axis) from SERP data that promoted distrust over the duration of data collection (X-axis) from October 5 to December 3, 2020. Content in headlines of SERP videos promoted increasingly more distrust than SERP stories in the post-election period after November 3, 2020.

Overall, though the focus in the pre-election period was primarily on the role of trust-undermining advertisements Zeng et al. (2021), videos appear to have been a far more challenging issue for content that cast doubt on the election's legitimacy. Moreover, given that videos are more difficult to monitor due to the challenges associated with tracking in-video content and graphics (Nakov et al., 2021; Jalli, 2021; Bradshaw et al., 2020), we believe that videos could be more delegitimizing beyond these headline differentials. For example, our data included videos with titles such as "LIVE 2020 Presidential Election Results," which, though coded as *Provides Information*, was found to be projecting false election results. Further research is needed to determine the scope of the use of misleading headlines to mask controversial in-video content and to capture deliberate efforts to evade censoring through the deployment of innocuous headlines (Moran et al., 2022).

4.5.2 R2: Geographic location

In our analysis of geographic trends, our subset of election-related Google headlines provides evidence of both effective stewardship and concerning patterns of distribution of delegitimizing political content. Our coding—to our surprise—did not identify any differences in the kind of content based on any combination of the "Stance" and "Promotion" categories that was served to cities based on their sizes (i.e., whether we classified the city or location as an urban area, urban cluster, or rural area as specified in Table 4.2). We suspect this to have happened because search engine platforms may not find it useful to personalize the results for smaller regions with a population of a few thousand people, especially when the news involves topics about the national election.

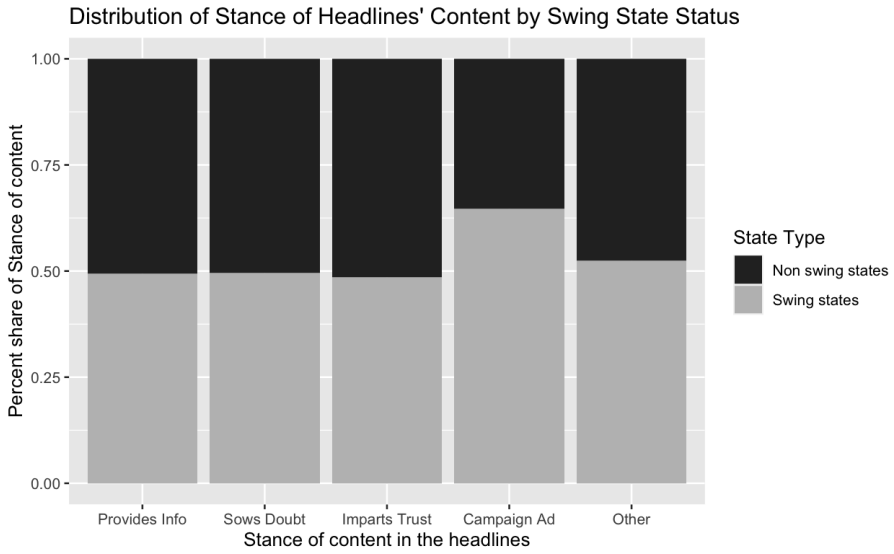


Figure 4.5: Searches made from swing states (total 6 locations in our collection) during the 2020 presidential election returned relatively higher percent share of campaign-based advertisements as compared to searches made from non-swing states (total 14 in our collection).

One notable difference between the swing states and non-swing states was the amount of campaign ads that appeared in the search engine results page. A multinomial regression analysis indicated that the odds of a campaign ad (compared to merely providing information) occurring in a swing state was almost twice that of a non-swing state. Figure 4.5 illustrates how these campaign ads almost always occurred through the advertising—and hence paid—SERP vertical in swing states as opposed to non-swing states. We found a similar pattern when investigating the difference across the electoral vote, with red states having more

campaign ads than blue states.

4.5.3 R3: Search terms

Moreover, while differences in political content were small across cities, focusing on conspiratorial search terms often led to politically biased and more frequent misleading search results. As previously noted in Table 4.2, our search terms consisted of two groups—one that focused on ordinary election terms and another that focused on conspiratorial topics. By using these two types of election terms as predictors, our models suggested that headlines containing content that promoted distrust in the election were about six times more likely to appear on Google SERPs when a user actively searched for a conspiratorial topic than when compared to more general searches about election topics (Table 4.6). Figure 4.6 shows the number of trust-undermining headlines that appear on the results page of Google Search given the various search terms inspected in this study. Headlines promoting distrust in the election increased considerably when we conducted searches based on conspiratorial terms.

Search term type	Odds ratio	CI [95%]	p-value
Sowing doubt and promoting it (Yes, No; reference: No)			
(Intercept)	0.252^*	[0.229, 0.276]	$< .001$
General search term	0.167^*	[0.135, 0.206]	$< .001$

Table 4.6: Odds ratios for "Sowing doubt and promoting it' when searching for general election-related terms as compared to conspiratorial election-related terms (described in Table 4.1) calculated using logistic regression.

Searching for specific terms during the election period did return content that promoted distrust in the election, but the rates were much higher for the conspiratorial terms. That is, individuals who sought out narratives that discussed potential issues with the election were not always directed away from delegitimizing content. This is not surprising, given that Google's business model emphasizes its ability to deliver the content most likely of interest to end users. However, it does place more emphasis on the process by which this content is selected and delivered (e.g., tagging, labels, etc.). For individuals searching for general election terms and questions, which likely included a far greater share of Google's users,[7] our data suggests

[7]According to Google Trends, only one query, "Newsmax election results"—which we believe might have displayed some delegitimizing content—appeared in the top 25 rising search queries on Google's search engine in the same time period as our collection; most other queries involved phrases like "election results," "Presidential election," "where do I vote," or "who is winning," which resonate with the general search terms that we used.

these users were subjected to fewer headlines containing content with the potential to undermine their trust in the election. Given this distinction, we see this as some evidence of successful limitation of pathways to delegitimizing content.

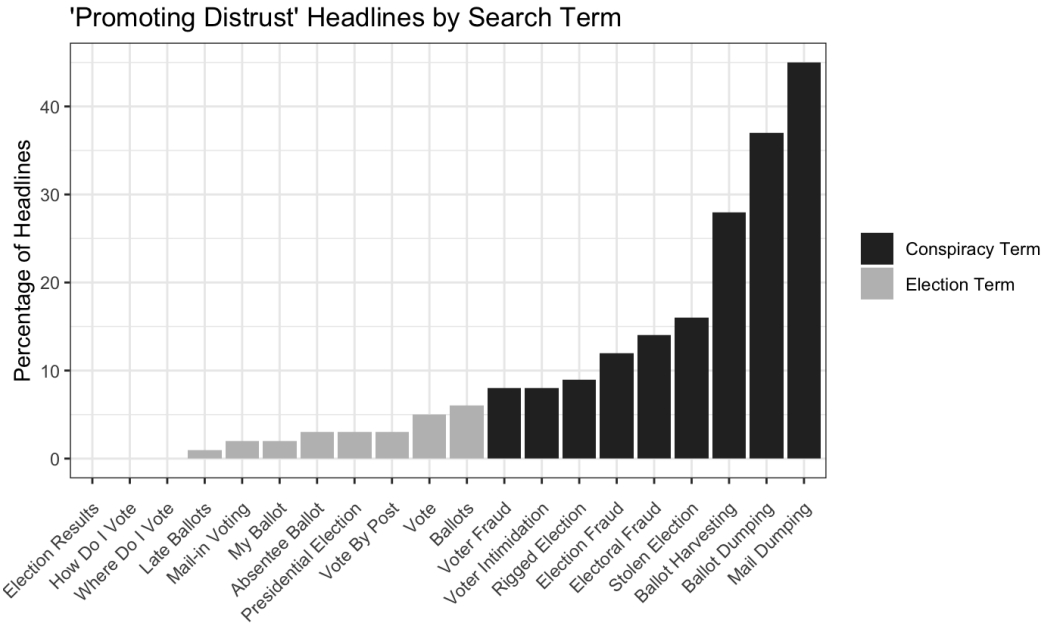


Figure 4.6: Frequency of doubt-sowing headlines given various search terms. Conspiratorial search terms that actively look for election-related issues served more delegitimizing content than the general search terms.

4.5.4 R4: Media domains

In addition to differences across SERP vertical type and location, our coding also revealed differences in the presentation of content across distinct news domains—with a specific focus on partisan outlets. We employed the media bias and media reliability scores from Ad Fontes Media (version 6.0) (Ad Fontes Media, Inc., 2020)—a choice based on recent research that also needed interpreting media bias of online news sources (Huszár et al., 2022; Brooks and Porter, 2020; Baranauskas, 2022; Zhao et al., 2020)—as predictors for investigating the effect of media partisanship on the kind of content served by these domains. As per these measures, a bias-score of +21.29 for OANN and -18.12 for Democracy Now! indicated partisan-right and partisan-left, respectively, in our data.

Consistent with expectations, our models indicated that with every unit increase in the bias of a media domain (implying higher right-leaning bias), the likelihood of a headline's content that challenges the integrity of election by sowing doubt (relative to mere providing information) increased significantly, by roughly 5.3% (Table 4.7). This trend continued when we accounted for how some media sources promoted the headlines that served to delegitimize the election's integrity; every unit increase in the bias scores of a media source (i.e., increasing right-leaning bias) could result in 2.6% higher chance of it promoting content with the potential to undermine trust in the election and about 4% lower chance of promoting content that reinstated public trust in the election (Table 4.8). Our models indicated no such effect for media reliability scores. Although Ad Fontes Media v6.0 data only accounts for 44% of the unique headlines from our sampled data, results indicate the severity of damage that partisan media could cause—by promoting debunked content in mainstream information channels like search engines—to public faith in democratic processes.

Type of stance	Odds ratio	CI [95%]	p-value
Campaign Ad (Intercept)	0.005^*	[0.006, 0.006]	$< .001$
Campaign Ad	0.877	[0.674, 1.142]	.331
Imparts trust (Intercept)	0.454	[0.139, 1.485]	.191
Imparts trust	1.002	[0.983, 1.022]	.829
Sows doubt (Intercept)	2.476	[0.931, 6.585]	.069
Sows doubt	1.053^*	[1.036, 1.071]	$< .001$
Other (Intercept)	0.031^*	[0.001, 0.341]	.004
Other	1.016	[0.977, 1.057]	.421

Table 4.7: Odds ratios for the different "Stance" type (relative to *Providing information*) reported for every unit increase in the media bias score taken from Ad Fontes Media v6.0 (Ad Fontes Media, Inc., 2020), calculated using logistic regression.

	Odds ratio	CI [95%]	p-value
Sowing distrust and promoting it (Yes, No; reference: No)			
(Intercept)	0.259^*	[0.079, 0.844]	.025
Bias score	1.026^*	[1.005, 1.048]	.014
Imparting trust and promoting it (Yes, No; reference: No)			
(Intercept)	0.133^*	[0.026, 0.674]	$< .015$
Bias score	0.961^*	[0.935, 0.989]	$< .005$

Table 4.8: Odds ratios for "Sowing doubt and promoting it" and "Imparting trust and promoting it" with every unit increase in the media bias score taken from Ad Fontes Media v6.0 (Ad Fontes Media, Inc., 2020), calculated using logistic regression.

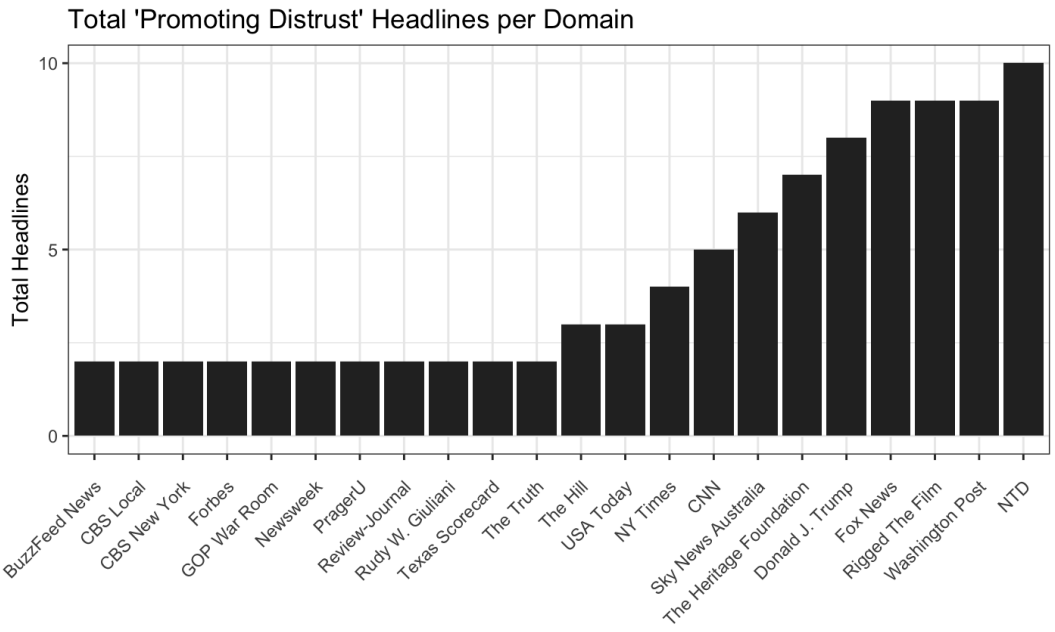


Figure 4.7: Media domains (X-axis) that promoted the most number of headlines (Y-axis) with delegitimizing content.

Moreover, many domains associated with content that promoted narratives with the potential to undermine electoral integrity with the highest frequency were affiliated with hyper-partisan outlets when examined both by the total and frequency of concerning posts. Looking first at total headlines coded as promoting doubt, [8] we find that while this included less reputable sites such as the Chinese-language site NTD (see Figure 4.7), the list also included activist organizations such as Rigged, which, though perhaps well-intentioned, promoted ads with headlines that served to undermine trust in the election. [9] More alarmingly, several prominent legacy news sites, including CNN and Fox News, also rank toward the top of total articles with these dual designations.

However, when looking at percentages (Figure 4.8) rather than totals for sites like CNN and Fox News, their comparable rankings stand out less. While this is encouraging, taken together these outputs serve as a reminder that due to the much greater quantity of information put out by many legacy news organizations,

[8]For both total and frequency calculations, only domains with more than three headlines appearing in the coding dataset were included in the plots.

[9]Users of the Google search engine were shown advertisements titled “The Voter Suppression Playbook - Watch ‘Rigged’ for Free” when searching for the keywords “rigged election” or “stolen election,” and when clicking directed to the URL https://www.riggedthefilm.com/

even a small share of concerning articles can play an outsized influence in delivering content with the potential to undermine trust in the election to the public.

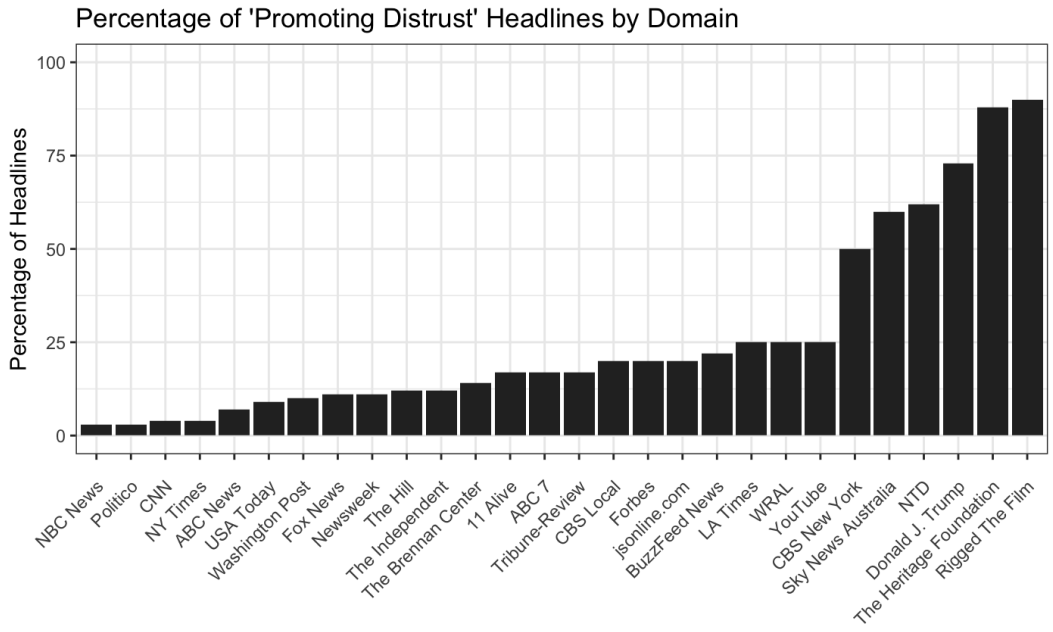


Figure 4.8: Percent share of unique headlines (Y-axis) per media domain (X-axis) that promoted delegitimizing content.

Based on a review of the headlines associated with the organizations that had the highest percentage of promoting doubt designations, emphasis on electoral fraud appeared to be the most common strategy to challenge the election's integrity and/or validity. In all, our preliminary foray into domain analysis should serve to initiate further examination of the sources of content with the potential to undermine trust in the election across both legacy and partisan media outlets. As with our media bias analyses, the narrow scope of our work here should serve only to draw broad conclusions regarding the types of headlines deployed across distinct forms of media groups rather than to identify specific domains for critique.

4.6 Discussion

4.6.1 Reflecting on Google Search as a socio-technically opaque gateway to the 2020 US election

Through this research, we found that Google SERPs gave access to concerning information provenances, but primarily when the search context involved conspiratorial terms or when information was accessed through the *videos* SERP vertical. However, for searches based on general election terms, Google did a relatively good job of surfacing relevant content without leading users to misleading arguments that negatively impacted civic trust in the election processes. Given the diversity of public opinion—sometimes at odds—across different regions in the United States, it can be challenging to deliver information that caters to the public interest yet steer clear of any regional biases. We were pleased to find no evidence that the search engine created information bubbles catering to any regional bias. The proportion of trust-undermining to trust-imparting content served in swing states was also similar to the proportion in non-swing states. We believe that Google offers at least some customization of SERP general results based on one's location to ensure such relative indifference to users' location when using the search engine.

Several researchers have looked into advertisements as a medium of information that serves misleading content—political ads, in particular, (Zeng et al., 2021; Kreiss and McGregor, 2019). While we found that campaign-based ads occurred more in our data, these were mostly placed by activism-based organizations, such as the ACLU [10] and Winred.com. [11] These ads did not seem harmful to perceptions of election integrity. Some other ads that our coding suggested had misleading content in their headlines occurred evenly across geographic locations.[12]

Though ads were comparatively less of a concern in our audit, video headlines served information provenances that consisted of information with the potential to undermine trust.[13]. Given that videos are more difficult for fact-checkers to check for misleading perspectives, it might bypass their scrutiny, compared

[10]The American Civil Liberties Union (aclu.org) is a nonprofit organization to safeguard human rights and liberties.

[11]Winred is the official online fundraising platform supporting the GOP.

[12]We identified a couple of advertisements in September 2020—prior to the period of data collection that we analyzed within the scope of this paper—that we believed contained delegitimizing content. These ads were taken down soon after we reported them to Google. We suspect that including these ads in the collection might have impacted the reported findings.

[13]Given the possibility that Google tends to overwrite the headlines of about 33.4% of the SERP content (Pecanek, 2021), it poses a question of whether Google's rewriting could play any role in altering the trust-undermining or trust-imparting potential of SERP headlines.

to the text modality. We did not have access to video usage, but it could be that videos also have more views. Prior research already points out that people tend to believe more in the information they see in a video than what they read through text (Wittenberg et al., 2021). In addition, plenty of studies have shown that clickbait is successful for a reason—people click on it (Scacco and Muddiman, 2016; Bhowmik et al., 2019). At present, researchers believe that Google's tools for stopping video-based misinformation seen on the YouTube platform are only partially effective (Hussein et al., 2020; Donzelli et al., 2018). With the possibility of Google including more videos from a broader range of platforms like TikTok and Instagram on its results pages, more work will be needed to monitor the impact of making "influencer" content easily accessible through search engine provenances. This is of particular concern given creators' ability to exploit the difficulties of monitoring videos and disguise content by evoking innocuous headlines, as occurred during the 2020 US election (Tenbarge, 2020). Our current audit is unable to track these additional issues.

In terms of actionable responses, though we acknowledge the challenges associated with video content management, we find that simple headlines can narrow auditors' focus onto concerning content without necessitating the designation of resources necessary to sample all election-related videos randomly. While this does not solve the issue of videos using deliberately vague or misleading headlines to hide controversial content, it could be used to limit the mainstream influence of similar videos by ensuring that they remain in the periphery without showing in the results page when users search for general election concepts, terms, and questions. However, nothing in our audit suggests that censorship should be promoted as a central strategy of search engines in managing political and politically adjacent content.

4.6.2 Designing future election-based audits

The included analyses were enabled by the strategic collection of data around the 2020 US election. Future analyses can build on these results in several ways.

First, with additional resources, we can refine the coding scheme and build it out to address a broader range of issues, topics, and concepts. Although we developed a rigorous coding scheme to make sense of the news headlines, we utilized only those codes in this research that focus on headline content with the potential to undermine trust as a compromise that allows for a longitudinal peek under the curtain while keeping the work manageable. With the collaborative efforts of the search engines themselves, it may be

possible to capture and categorize similar headlines in real-time and match headlines with the associated content of each SERP vertical type. For instance, pairing potentially undermining headlines with the nature of the underlying video content might generate a more nuanced understanding of the pathways connecting users to political content and generate knowledge of how these components intersect. Further investment in post hoc coding may also enable differentiation between types of potentially undermining content, such as misleading content and outright false content.

Second, future audits can inform the priorities of search engine staff during election periods. While Washington, DC, and Silicon Valley have given much attention to the content linked to political advertisements, our audit suggests that when compared to other modalities, advertisements may not constitute the primary provenance from which users encounter content with the potential to undermine electoral trust. Further research into the different sources of content promotion should allow search engines to allocate resources more efficiently across their networks.

Third, we recommend that auditing reports should be thorough, comprehensible, and easily accessible to different stakeholders so they can contribute in meaningful ways to safeguarding the trust in election processes. For instance, although Google publishes a list of the political ads hosted on the search engine as the "Google transparency report on political advertising," the vague criteria of what constitutes a political ad and the limited information it requests about an ad publisher make it easy to circumvent the report's scope. For example, our extended data collection contained ads from protectthevote.org (paid for by the Republican National Committee) that appeared in the transparency report, but ads from "protectmyvote.org" did not seem to fit Google's criteria of political advertising.[14] In addition, platforms should make search engine data available to researchers so they can serve as independent third-party auditors and help monitor the health of these information provenances as they vary for individuals based on different search characteristics. This research was possible because we paid a third party for the API access, which is a financial hurdle and a caveat that might impact the quality of the data. We acknowledge that when collecting data through different sources, it is important to protect people's privacy and believe that the data we accessed poses less threat to user privacy than SERP data collected through browser plugins (e.g., (Robertson et al., 2018). We believe that law should require search engine platforms to provide researchers with access to anonymized data, as

[14] As reported in the article (Stanley-Becker, 2020), Google took five days before they removed the "protectmyvote.org" ads from their platform after its discovery.

nothing in Section 230 or the First Amendment stands in the way of such transparency.

Previous audits on search engines like Google Search have discovered several insights into how these platforms can shape public opinion, especially around critical topics like elections (Hu et al., 2019; Mustafaraj et al., 2020; Trielli and Diakopoulos, 2019, 2022; Diakopoulos et al., 2018; Robertson et al., 2018). Our research adds to this body of literature on how Google fared in delivering election-related news to its users across America in 2020. In addition, we curate a dataset consisting of 47,000 advertisements, 500,000 main search results, 240,000 news stories, and 66,000 videos—each of which constitutes a unique information provenance—and make it publicly available (Zade et al., 2022a) to facilitate the discovery of more insights about the 2020 US elections as pictured through the agency of search engines.

4.6.3 Conclusion: Towards contextualized information provenance

In this Chapter, our analysis demonstrated that controversial content did not surface on the Google Search platform when users searched for more general keywords like "election results" or "presidential election." More encouragingly, we found that during the 2020 US election, individuals who searched for general election terms, issues, and questions—which we believe to be the dominant set of users—were largely shielded from headlines that could undermine electoral trust.

Unfortunately, the Google Search platform offered some access to controversial information provenances when the search context involved specific characteristics. For example, headlines like "The biggest election fraud story you haven't heard about..." only showed up when users searched for the keyword "election fraud" that has a conspiratorial context. This can be concerning given that SERP headlines are known to offer its users more partisan cues as compared to the original webpage (Hu et al., 2019). This remains a tricky area rife with opportunities for more research to determine how to serve users when their search actively involves characteristics that might give them access to information provenances involving delegitimizing content. Are there ways to add more context about the provenances to aid users in making better judgments about assessing the credibility of information they access through those provenances? Given the recent update about the Google Search platform allowing its users to add search notes (Peters, 2023) — similar to Birdwatch notes on Twitter (Coleman, 2021) — it is important to study how adding context to information provenances might impact credibility assessment of the information by users. In the next Chap-

ter, I address this opportunity as I examine the potential of serving users with contextual cues consisting of trajectory and account history on the Twitter platform.

Chapter 5

Interpreting information opacity on Twitter: Investigating trajectory-based cues to contextualize how information propagates

This chapter discusses the research I spearheaded and had published in the Proceedings of the ACM on Human-Computer Interaction, volume 7, in 2023 (Zade et al., 2023). Throughout this chapter, the pronoun 'we' encompasses all the co-authors involved in the project. I have incorporated substantial portions of this chapter from the published work, with adjustments made to the introduction and background sections, as well as the design of the intervention section. Additionally, I have omitted certain parts of the discussion from this chapter and reworked them within the broader context of the overall discussion presented in Chapter 7.

5.1 Introduction

Well-intentioned users sometimes enable the spread of misinformation due to limited context about where the information originated and/or why it is spreading. The Covington Catholic High School controversy

(introduced in Chapter 1 is one good example of such misinformation that went viral due to missing context about a faceoff between a Native American man and a group of high school boys and received about 2.5 Million views and more than 14k retweets (Wise, 2019). The video, which was originally posted on Instagram with incomplete context, was popularized on Twitter through the agency of an account that had misleading profile information and was subsequently suspended by Twitter (Eli, 2019). Another example involves the spread of misinformation during the 2020 presidential election campaign, claiming that protestors were transported to different events by one of the contenders without any evidence (Zadrozny and Colins, 2020). This false narrative suggested that these actions were orchestrated to manipulate the perception of widespread support or opposition, depending on the prevailing political narrative.

Though efforts to detect amplification by bot accounts on these platforms may be useful (Varol et al., 2017), 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 design of the retweeting scenario on Twitter 5.1 does not provide much insight (except the root, number of likes, and number of replies) about the spread of the tweet. Following recommendations from others (Institute, 2021; Cohen, 2021), here I explore the potential of supporting media literacy principles — access, create, analyze, evaluate and act upon information critically — through 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.

Figure 5.1: 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 help users identify misinformation, social media platforms—that only displayed popularity metrics

in their early days—started introducing warning labels Zhang et al. (2021a); Mena (2020); Lee (2020). 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 (Im et al., 2020). To advance the scholarship on social signals and help users identify if a Twitter account is acting in good faith or bad, in this paper, I propose that providing **Activity Cues** based on an account's recent online activity can provide insight into how that account operates and signal any potentially problematic behavior. Our chosen activity 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 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, I believe that these platforms should focus not only on the root (*i.e.*, who created the information) but also on the actors involved in diffusing the content. Inspired by the concept of a trail (Finn et al., 2015), and the importance of highlighting information provenance as realized in Chapter 4, I 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 [1]. 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) (Metzger et al., 2010), a popularizer account (that popularized the tweet) (Institute, 2021), and a friend account (that may have liked the tweet, leading it into the feed of the user) (Turcotte et al., 2015; Geeng et al., 2020).

For answering *RQ2*, *i.e.*, how can easy access to informational context — presented as a provenance — on a platform (Twitter) help users assess the credibility of that information, I adopted a research through design approach (Gaver, 2012; Zimmerman et al., 2007). I interviewed 21 participants—who were diverse in their political alignment and life experiences—using the intervention that I personalized to their individual tweeting experience and environment. I then analyzed the data using an affinity-diagramming-based

[1] Within this chapter, I use 'trajectory' to refer to the specific intervention used in this study. Provenance refers to the broader construct in which I situate the scholarship of this dissertation

thematic analysis. The contributions of this work include propose a design intervention modeled after the media literacy principles of accessing, analyzing, evaluating, creating and acting upon information critically to reimagine the retweeting scenario:

- By examining the concept of **activity-based cues**: I 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 (Klurfeld and Schneider, 2014; Fleming, 2014) 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**: I 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. I use the findings to argue why resurfacing the role of institutional credibility obscured by the networked nature of social media (Hancock, 2020; McCrosky, 2020) can be useful to curtail the spread of misinformation.

5.2 Background

5.2.1 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 (Hermida et al., 2012; Lipsman et al., 2012; Hill et al., 2017). 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 (Zhang et al., 2021a; Mena, 2020; Lee, 2020; Vraga et al., 2020; Koch et al., 2021). Preemptive labels aimed at inoculation have been effective to make people aware how they can be misinformed thereby increasing their resilience to it (Lewandowsky and Van Der Linden, 2021; Vraga and Bode, 2021; Roozenbeek and van der Linden, 2019). 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 (Bail et al., 2018; Peter and Koch, 2016; Swire-Thompson et al., 2020).

To support everyday platform users contribute and guide the application of these labels, in 2021 Twitter introduced a community-based approach called Birdwatch (Coleman, 2021; Pröllochs, 2022) so that users can add context to a tweet. Such labeling unfortunately has been found to support partisanship rather than promote fact-checking (Allen et al., 2022). 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 (Gupta et al., 2014).

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 (Kang, 2010; Donath, 2007; Gupta et al., 2014). For example, using labels that reflect accuracy of headlines have been found effective towards reducing the sharing of false information (Jahanbakhsh et al., 2021). Researchers have demonstrated that automatically derived nudges can be effective towards providing some context into the credibility of information (Im et al., 2020; Bhuiyan et al., 2021). 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 (Edgerly et al., 2020). In addition, these approaches are more suitable to the automatic way of thinking (over reflective) (Caraban et al., 2019; Hansen and Jespersen, 2013) to support quick decision making. To help identify information machinations of a more organized nature and preserve user agency, I believe that nudges that support reflective cues like the intervention that I propose can be more effective.

5.2.2 Information provenance, path and propagation

The source of information is considered to be one of the prime factors for assessing credibility of the content (Metzger et al., 2010; Jabiyev et al., 2021). Despite the challenges of identifying the information provenance and verifying its credibility for researchers and media platforms (Starbird et al., 2019; Diakopoulos et al., 2012; Figueira and Oliveira, 2017), cues about the source of information may not always be effective towards assessing content-credibility and detecting misinformation (Dias et al., 2020).

To benefit from other signals apart from information provenance, researchers have utilized diffusion-based metrics to help users assess credibility of the content (Marinova et al., 2020; Resnick et al., 2014). 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 (Finn et al., 2015; Metaxas, 2015). Shao introduced Hoaxy platform along similar lines to facilitate the collection, detection and analysis of all incoming tweets to detect misinformation online (Shao et al., 2016). 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, I introduce Twitter trajectories that illustrate the spread of information in a simpler fashion by highlighting specific actors involved in facilitating that content reaching the user consuming that information.

Researchers are aware that information spreads online through a network (Arif et al., 2018; Starbird, 2020). Network properties and behavioral features of propagation have been found to be useful for discerning misinformation from good information (Zhao et al., 2021; Molina et al., 2021). For example, 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 (Westerman et al., 2012). My approach exposes the behavioral traits and offers a much richer insight into the social network of multiple actors involved in putting information out there to provide more access to contextual information and aid its analysis by end users of media platforms.

5.3 Designing the Intervention

I adopted a research through design (RtD) approach for this research (Gaver, 2012; Zimmerman and Forlizzi, 2014). RtD is an approach to conduct research that facilitates discovery of new knowledge using methods of design practice (Zimmerman and Forlizzi, 2014). 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 (Carroll et al., 2020), including that of misinformation (Carroll and Bonkel, 2021; Sherman et al., 2021). Along similar lines, I 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 and context of different actors who are involved in spreading the information, I first came up with a framework that later guided our selection and design of the activity cues. Next, I 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. I now describe the process of designing the intervention.

5.3.1 Activity 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, I 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 problematic accounts or 'pandering for social capital' (Boyd et al., 2010); using hashtags that consist of hate can signal producing radicalized content (Agarwal and Sureka, 2015).

During the generative brainstorming process, I came up with 46 different (but not necessarily exclusive) cues that I 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, I intended to select a subset of these 46 cues such that they represent several problematic behaviors (as opposed to only one). Accordingly, I chose cues that characterized behaviors like amplifying problematic content, engaging in some organized inauthentic behavior, excessive activity towards a singular or specific few accounts etc.

With a focus on operationalizing the different cues based on the information publicly available on Twit-

ter, I along with my colleagues 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 variability in their subjective interpretation. I 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 others might use that hashtag on the platform. Similarly, the frequency of how often another account was retweeted by a Twitter account conveys the strength of shared beliefs between those two accounts (Boehmer and Tandoc, 2015).

Table 5.1 describes the different cues that I decided to include in the intervention. For each identified cue, Table 5.1 also indicates how the cue achieves the goals of media literacy principles to highlight what the design intervention personifies. It is possible that each cue can provide a signal into multiple aspects of media literacy, e.g., while trajectory provides access to previously hidden popularizer, it also sets the tone for asking questions about information purpose to help users analyze the propagation critically. Figure 5.2 provides more insight into these cues that I included in the intervention.

Figure 5.2: Representation of a cue card (center) that I showed to a participant. Each cue card is populated with activity 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.

Aligned with the recommendations of Aspen Institute's report 'Commission on information disorder'

Table 5.1: The different contextual cues that I included in the overall intervention: tweet trajectory and activity cues (based on the recent 4 weeks of account activity using Twitter API v2). I also indicate how each of the cue aligns with media literacy principles of accessing, analyzing, evaluating, creating and acting upon information critically.

Cues in the intervention based on four weeks of activity of an account	Corresponding **Media Literacy principle**
Tweet trajectory with a popularizer account	It provides users with a higher sense of *access* into the propagation of information which was otherwise not readily available, and help them *analyze* by probing about the possible purpose behind sharing that information.
Two most frequently used hashtags by that account	Hashtags serve an introduction into the kind of content an account might engage with online thus helping users *evaluate* the context.
One hashtag used by that account along with another account that popularized it	Helps users to *access* the account with whom the subject of the cue card might ideologically align, thereby providing richer context for a critical *evaluation* of the information.
Two most retweeted accounts along with the frequency of retweeting them	Provides an opportunity to *analyze* the immediate account affiliations and neighborhood of that information and also *evaluate* those affiliations subjectively.
Two most frequently shared domain names by that account	Provides an opportunity to *access* the account's regular media affiliations and *evaluate* them using one's own social-political understanding.
Frequency distribution of account's activity: tweets, quote tweets, retweets, & comments	Provides quick and easy *access* to historic data that shapes the further *analysis* of that account.
Distribution of other accounts by their Twitter age who retweeted this account	Generates new information that was previously unavailable to everyday users and makes it *accessible* to them.

— particularly about empowering users through digital interventions that give them the skills and context to safely navigate low quality (Institute, 2021) — I 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. I believe the proposed cues bring attention to the much needed human aspects of information propagation (Fernandez and Alani, 2018), *i.e.*, the personal motivations of why different accounts may post or share certain information.

5.3.2 Tweet trajectory

Finn et al. introduced the concept of Twitter trails that allows users to investigate how tweets that are potential rumors originate from a source and then propagate within a social network (Finn et al., 2015). Following up on the concept of a trail, I introduce a Tweet Trajectory 5.3 to demonstrate how the tweet could have reached the user's Twitter feed—with a particular focus on Twitter accounts that were responsible for bringing widespread attention to the tweet, *i.e.*, popularizers of that tweet.

Figure 5.3: The proposed "tweet-trajectory" consists of the root tweeter account, a popularizer account, and a friend account.

I chose our first actor to be the known source of the tweet (root tweeter) given the established significance of source for online credibility assessment (Metzger et al., 2010; Geeng et al., 2020). Realizing the potential of top accounts with large number of followers in spreading content (Institute, 2021; Chong and Kim, 2020), I 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 (Turcotte et al., 2015; Geeng et al., 2020), I 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 (Geeng et al., 2020).

By displaying the past activity-based cues in a trajectory, our intervention 5.4 aims to help users identify and understand specific manipulation tactics—which is a critical requirement of new media literacy (Saltz et al., 2020).

5.3.3 Reflecting on the design intervention

After identifying the activity cues of interest from publicly available Twitter information and selecting the actors for the tweet trajectory, I engaged in an iterative design process to create a prototype that balances the role, appearance, and implementation complexity, following principles outlined by Zimmerman (Zimmerman et al., 2007).

As part of this process, I conducted a user-research activity involving feedback collection from five graduate students enrolled in a University-based technology-design program. Initially, I presented them with three different styles for presenting each activity cue. Based on their feedback, I finalized the design choice that effectively communicated the intended information, as documented in Table 5.1. Subsequently, I presented the graduate students with two options — horizontal and vertical — for visualizing the tweet trajectory. Feedback indicated that the horizontal choice was suitable for the desktop version of the intervention (used in this study), while the vertical choice was better suited for the mobile version.

Based on the gathered feedback, we refined the design of the intervention, encompassing both the activity cues and the tweet trajectory. We then utilized this refined version as a probe in interview sessions. Although the findings of this research are closely tied to our design decisions, we believe our meticulous design process ensures that the findings remain robust even with minor design adjustments.

Given that a design intervention conveys a specific framing of the problem that one wishes to explore, I note that knowledge generated through RtD is reflective of the functions and limitations of our intervention (as understood by the participants) (Zimmerman et al., 2007) and as imposed by the researchers (Dow et al., 2013). To minimize bias stemming from specific visual design choices, we emphasized to study participants the importance of focusing on the information content rather than visual aspects like font size or color. This reminder was essential to ensure that participants' insights were based on the substance of the intervention rather than its visual presentation. Similarly, we took care to interpret interview data without being unduly influenced by the design of the cues themselves. Our goal was to draw conclusions based on the implications

of specific cues rather than their design features. I believe that the intervention we adopted in this research study is novel, highly relevant to providing informational context, and extensible for other researchers to build upon.

5.4 Study Design

5.4.1 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 trajectory-based intervention, I 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 trajectory-based intervention that I showed them.

I 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, I next posted the same recruitment survey on the Mturk platform (Paolacci et al., 2010). Upon realizing that most of the interest from the Mturk platform included participants without active Twitter accounts, I next posted the recruitment survey in the Craigslist-volunteers' section in several American cities that I adjusted according to the need for diversity of our participants.

I filtered 24 participants suitable for our requirement from about 90 expressions of interest that I gathered from the call for recruitment. I 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 about 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,

Table 5.2: 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

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 5.2 describes participant characteristics in detail.

5.4.2 Personalizing the intervention for each participant

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

Selecting the tweet: For every selected participant, I 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, I ensured that the participant did not follow the Root tweeter.

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

Selecting the friend: Once I selected the popularizers, I 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). I 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.

Populating cues: For every Twitter account in the trajectory (root, popularizers, friend), I populated the respective cue card based on their publicly available Twitter profile information and their publicly accessible Twitter activity as indicated in 5.2.

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

5.4.3 Interview procedure

Each interview lasted about 60 minutes and involved one participant, one researcher to guide the discussion, and one researcher to take notes. I conducted the session online over Zoom and compensated the participants

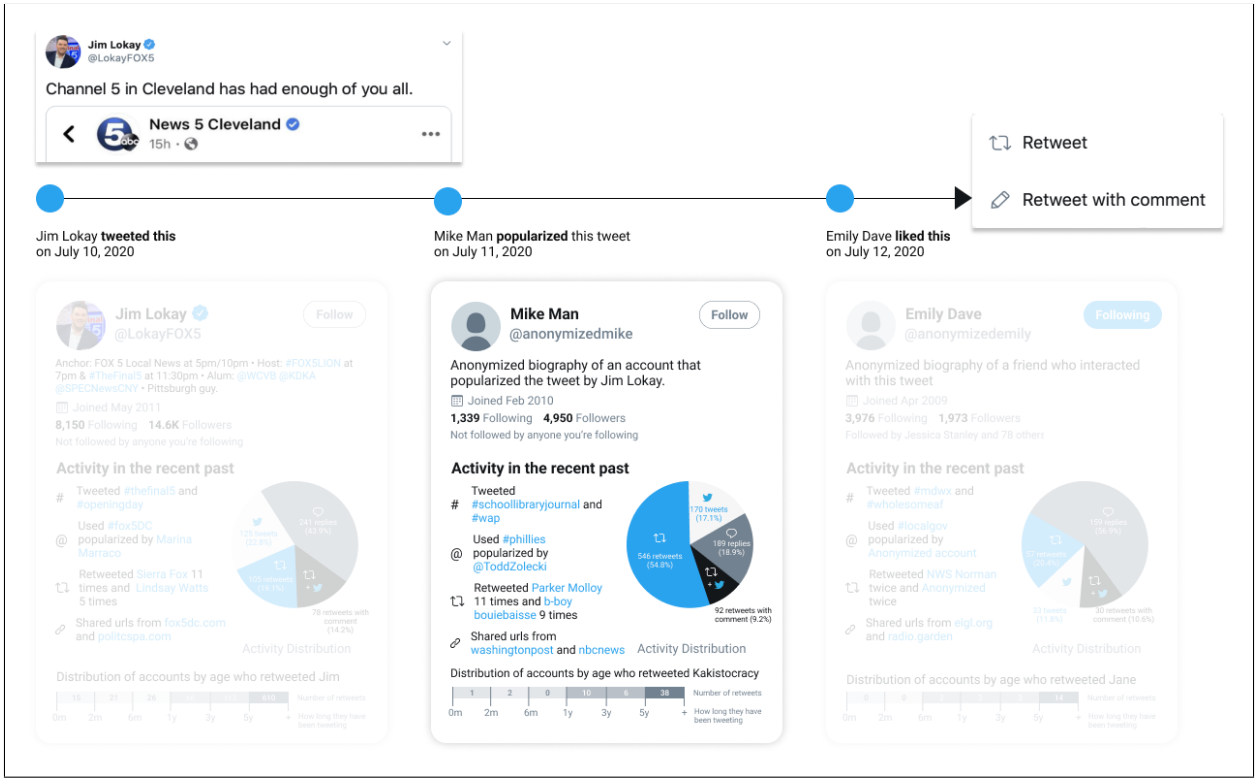


Figure 5.4: Reimagined retweeting scenario using activity cues and tweet trajectory corresponding to Jim Lokay's tweet personalized for a study participant. Jim Lokay here represents the root, Mike Man represents the popularizer, and Emily Dave represents the friend who may have liked and brought this 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.

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, I 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. I 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 5.2) in four steps. In each step, I 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. I 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, I showed them the tweet that they had retweeted in the past and that I chose to populate the cues in the personalized intervention. I 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. I note that some of the participants were not too certain if they had shared the tweet citing that retweeting is a snap decision.

Next, I showed them one of the trajectories of that tweet as illustrated in Figure 5.3 and asked them what the illustration means to them. I then showed them our designed trajectory conceptualization (Figure 5.4) and asked them to talk aloud about how they might use (if at all) the information in the intervention to guide their decision-making. I 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.

5.4.4 Data analysis

I 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, I 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. I iteratively refined these themes and recorded the insights using analytical memos. In section 5.5, I report on how activity cues support participants in developing and/or refining their mental model about an account. In section 5.5, I then report insights about how the tweet trajectory complements these efforts towards making a quick-but-informed credibility judgment.

5.5 Summary of Findings

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 media literacy principles, 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. 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.

> "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

The trajectory part of the intervention (Figure 5.4) 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 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. I 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.

Table 5.3 summarizes the findings from studying media literacy-driven contextual cues regarding Twitter users' retweeting experiences. I discover several problematic behaviors and underlying causes of concerns corresponding to each of those behaviors. The complete set of findings is available here in this published research.

Table 5.3: Activity 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	Questionable **purpose**	Self-described or user-inferred propaganda account
	Inconsistent **participation**	Journalist account with no original tweets
CC2: Online associations of an account	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	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	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	Potentially create **problematic association**	Retweeting creates new cues (online activity) that are not representative of account's history and/or purpose

5.6 Discussion

Information that I come across online in everyday life can be highly contextual (Eli, 2019). 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, I 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 their users realize how different intentions of those who share information might shape the larger context around it and compromise its credibility. I address this missed opportunity through this research and introduce the media literacy-based intervention consisting of activity 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. I now briefly introduce the limitations and then discuss the findings (summarized in Table 5.3) on 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.

5.6.1 Limitations

Given the limitations of Twitter API, I could only access the 100 most recent retweets. As a result, I cannot say with certainty if the popularizer, whom I showed in the intervention, is a true popularizer for that tweet. Another limitation of the intervention is that I 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, I 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. Further, I suspect that some participant selection bias might impact our findings. Though I tried our best to recruit diverse participants with diverse life experiences and political ideologies, their willingness to participate in such a study can indicate their tendency to be critical of misinformation in general.

5.6.2 Contextual cues help assess overall credibility of an account

The first part of the intervention involved activity cues that signal intend to signal different inappropriate behaviors that account exhibit online. Using research through design as our method, I discovered that our activity cues helped support users in assessing credibility in three ways. First, it 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 create 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.

The activity 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 associates with 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 (Klurfeld and Schneider, 2014; Fleming, 2014). 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, the activity 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 inauthentic organized behavior. The activity cues thus enabled users to assess an account's credibility by employing context—not readily available on Twitter— the cue cards offered that.

5.6.3 Provenance helps assess the credibility of propagating information

Users employed the easy access to popularizer accounts given by the trajectory part of the intervention — provenance by extension — towards aiding their judgments. First, the trajectory made users evaluate their trust—an important factor for credibility assessment (Kim and Brown, 2015; Sternthal et al., 1978)—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 before 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 allowed users 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 (Cheng et al., 2017; Grabner-Kräuter and Bitter, 2015) in a way that is not facilitated by the present design of Twitter.

At times, trajectories surfaced tension as users found the information 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 based on rewards and penalties (Grabner-Kräuter and Bitter, 2015).

The present design of social media—as Hancock describes (Hancock, 2020)—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 (McCrosky, 2020). In other words, users tend to trust low-quality content shared by their trusted friend more than high-quality content shared by distant friend Sterret et al. (2018). I 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.

My proposed choice, *i.e.*, the design of the intervention and the selected cues in it is one of the many ways of designing such a provenance-based intervention. It is possible that 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. I note that such a variation does not compromise the validity of the findings but is reflective of the research through design method (Gaver, 2012) adopted in this research.

The intervention provided users information that—though publicly available on Twitter—users had to seek out and synthesize, which is not feasible within the retweeting scenario. Future iterations of the intervention can incorporate the principles of lateral reading technique (Wineburg and McGrew, 2017) to enable quick assessment of multiple actors in the provenance-based intervention and further promote platform-driven media literacy.

Chapter 6

Signaling information opacity on TikTok: Identifying contextual cues that translate to other platforms

This chapter delves into my original research conducted on the TikTok platform, which remains unpublished. Throughout this chapter, the pronoun 'we' refers to all the co-authors — Zade, Himanshu and Choi, Sehe J. and Mai, Huy and White, Emily and Favro, Sophie and Vasquez, Keiver Bencomo and Khor, Tyler and Turns, Jennifer — who contributed to the project. The content presented in this chapter is entirely new and is being published for the first time within this dissertation.

6.1 Introduction

In the previous Chapter 5, I studied provenance-based contextual cues to help Twitter users understand how misinformation could propagate on Twitter to their feeds. I discovered several problematic behaviors exhibited by Twitter accounts that users need to identify for assessing credibility of information and mitigating the propagation of misinformation. For example, one of these behaviors included examining if an online account has a questionable purpose or if that account shared any content from an uncredible media source. In this chapter, I introduce another study that identifies the list of similar problematic behaviors that users

need to identify in the context of TikTok platform.

TikTok has several unique features that have led to its massive adoption across the world Herrman (2019). These include the "For You" page with algorithmically created content catered towards individual users (Anderson, 2020), the "Duet" feature that splits the screen in two to simultaneously display comparative video content, etc. Given the prominence of viral online trends that could influence everyday habits on TikTok (Kriegel et al., 2021; Korbani and LaBrie, 2021), it is critical to understand what processes users adopt to discern the harmful content from harmless content. In this chapter, I introduce my study on how users employed the TikTok design features to assess content credibility, particularly when considering sharing the content (or not). In other words, what TikTok affordances — that serve as signals of problematic behaviors — help users assess content credibility on TikTok?

The algorithmically curated TikTok feed, sometimes containing political information, is mediated through videos. Though video content may not be as persuasive as was once feared, research indicates that users tend to believe the information in a video more easily than that in text format (Wittenberg et al., 2021). Given the increased prevalence of video-based misinformation — *e.g.*, 20–32% of Covid-related videos between January and March on TikTok contained misinformation about the pandemic (Southwick et al., 2021) — there is a shared belief among researchers that the real extent of persuasiveness of videos might diverge in real settings that are not lab-controlled. In addition, when comparing the role of text versus video modality within messaging apps, researchers found that users process videos superficially and tend to be more influenced by them compared to text (Sundar et al., 2021).

Videos are also more difficult to monitor due to the challenges associated with tracking in-video content and graphics (Nakov et al., 2021; Jalli, 2021; Bradshaw et al., 2020). Research indicates that though debunking can only marginally aid people's ability to discern false videos, their belief in the misinformation claim remains significantly lower in the debunking condition compared to the misinformation-only condition (Bhargava et al., 2023). My research extends on this promise to study TikTok-specific affordances that further aid people to discern misinformation.

To answer *RQ3*, i.e., how users employ the TikTok platform cues to assess information credibility, I conducted participatory design research in which, first, researchers served as participants and documented their experiences of using TikTok affordances to assess the information credibility of videos related to ten

selected topics from two accounts varying in political affinity; second, researchers clustered and organized these experiences to identify how, and which, cues users employ when assessing video credibility on TikTok. Contributions of this Chapter include:

1. Demonstrating that TikTok-specific cues can effectively signal socio-technical dimensions of information on the TikTok platform to its users. I highlight how the concept of information provenance becomes useful in the process of credibility assessment.

2. Identifying the unique processes of assessing credibility on TikTok. I focus on the collaborative processes that users adopt to discern misinformation and challenge TikTok's algorithmic mediation.

6.2 Related Work

6.2.1 Focus on algorithmic curation of online feeds

Like other online platforms, the TikTok platform uses algorithms to suggest posts to its users. The default video feed on TikTok is the algorithm-driven "For You" section; however, users may exercise the option to choose to limit their video feed to just those accounts they follow. This focus on personalized content discovery through the "For You" page has led TikTok to be commonly seen as more driven by algorithms (Anderson, 2020). On the other hand, Twitter (at least till 2022) struck a better balance by giving users the option to choose between chronologically curated feeds and algorithmically curated feeds. This distinction affords TikTok a higher potential to push viral content to users.

Feeds primarily curated with algorithms are highly susceptible to the challenges of algorithmic amplification, *i.e.*, the process of recommending certain information to users with a primary intent of increasing user engagement (Menczer, 2021). For example, such amplification can lead to more resilient echo chambers that strengthen conspiratorial beliefs (Peck, 2020; Grandinetti and Bruinsma, 2023; Forberg, 2022), extremist content (Whittaker et al., 2021), political polarization (Barberá, 2020), etc. This might also amplify the individual biases in some instances as one group might benefit more than the others (Huszár et al., 2022). With increased access to viral and algorithmically driven content, TikTok passes the onus of ascertaining content credibility primarily on users as an individual's responsibility (Li et al., 2023).

6.2.2 Online trends and influencer culture

Instagram was one of the early platforms to allow individuals to share visual content, build their following, and monetize their influence (Hearn and Schoenhoff, 2015; Abidin, 2016). These individuals, now referred to as "influencers," are known to impact the purchasing decisions of their followers (Młodkowska et al., 2019; Wulandari and Darma, 2020). This influence doesn't merely stop at marketing but also extends to opinions and topics of importance, including politics and science (Riedl et al., 2021; Bause, 2021; Naderer, 2023; Chinn et al., 2023). Although the captivating appeal of science highlighted in TikTok memes attracts the attention of young people, unfortunately, it could also reflect an uncritical approach to scientific concepts; *e.g.*, researchers found that TikTok memes have the potential to depict science as an individual pursuit rather than a collaborative endeavor (Zeng et al., 2020). In science-related news, highlighting prominent individuals rather than scientific communities or the broader scientific system has held significant news value (Badenschier and Wormer, 2011). Emphasizing on identities of prominent individuals has often drove political affective polarization (Iyengar et al., 2019; Gill, 2022). In the TikTok age, influencers could serve that prominent person's role as they generate viewership for content in a potentially uncritical manner (Zeng et al., 2020). Given such a potentially frivolous interpretation of science, it remains critical to understand how users of TikTok engage with the design features to afford themselves a better assessment of the credibility of that information.

6.2.3 Demographics of TikTok users

TikTok has been particularly popular among younger audiences. Though the platform has exhibited growth among users in their 30s, a significant portion of users falls within the age group of 18 to 29; Pew Research Center found that a majority of 18 to 29-year-olds use Instagram (71%) or Snapchat (65%), and approximately half used TikTok (Auxier and Anderson, 2021). Having a relatively younger group of users makes TikTok different from Twitter.

Younger Americans, often called digital natives (Prensky, 2001), have grown up in an era where digital technology, the internet, and social media are prevalent. This generation is more familiar with navigating online platforms, fact-checking information, and discerning between different media sources (Helsper and Eynon, 2010; Coskun, 2021). Younger generations may have had more exposure to media literacy education

as part of their formal schooling, as educational institutions inculcated media literacy education in the school curriculum, equipping them with skills to evaluate information critically. How can the newer generation utilize this strength to make sense of the new affordances within the platform? I use this opportunity to investigate if young TikTok users, when assessing content credibility, employ information provenance as a cue of credibility; if so, in what unique manner does the employment occur? I now discuss the methodical setup to benefit from such a choice of study participant sample.

6.3 Study Setup

6.3.1 Methodical Approach

For this study, I adopted a Participatory Research (PR) approach, *i.e.*, a research-to-action approach that emphasizes direct engagement of local priorities and perspectives (Cornwall and Jewkes, 1995). This approach prioritizes co-constructing research through partnerships between researchers and stakeholders, community members, or others with insider knowledge and lived expertise (Jagosh et al., 2012). Accordingly, this research is built upon the lived experiences of TikTok users who contributed to the data and as co-researchers who then analyzed that data.

6.3.2 Participants

Following the participatory research approach, this research is built upon the lived experiences of TikTok users. The study participants, whom we will refer to as co-researchers, were recruited as a part of a research project conducted by a lead researcher enrolled in the Human-Centered Design and Engineering department at the University of Washington. All the co-researchers had prior experience of at least six months with the TikTok platform and received academic credits towards participating in this research.

6.3.3 Data

Initial Setup of TikTok accounts

For a period of four weeks, participants collected data, *i.e.*, their impression of how engaged with a TikTok video that they watched about specific search terms provided to them (refer to Table 6.1). To ensure that

the search was being made from TikTok accounts that were representative of the political diversity of users, we adopted the following process to set half of the accounts as politically left-leaning and the other half as politically right-leaning:

1. First, every participant-researcher set up 2 TikTok accounts for your own use.
2. They read the research by Lund et al. that studied the role of TikTok's algorithm for promoting polarization and found that the algorithm curated different content to a liberal user, a conservative user, and an independent user (Lund and Zhong, 2018).
3. Next, participant-researcher focused on the table in that same research paper (Lund and Zhong, 2018) — can be referred to in Figure 6.1 — that documents the variation in hashtags that shown up across the three kinds of users. We used these hashtags to seed any kind of political leaning into the accounts that we later used to conduct this research.
4. For the first of the two accounts, every participant-researcher chose any hashtag from the "Liberal tags" column in Figure 6.1. They followed any account that showed up on the first screen without scrolling to take advantage of TikTok's recommended accounts.
5. Participant-researchers repeated the above step nine more times to set up the first TikTok account with ten left-leaning hashtags.
6. Participant-researchers repeated steps four and five above for their personal second account with "conservative tags" as seen in Figure 6.1.

Daily logs of TikTok-video interaction

Upon confirming that each participant-researcher had set both of their TikTok accounts — first with ten left-leaning hashtags and the second with ten right-leaning hashtags — the lead researcher then guided all the participant-researcher to adopt the following procedure to document the experience of how every participant-researcher engaged in watching a TikTok video with a particular focus on how they employ TikTok design features available to them to assess credibility of the video:

1. Go to one of the Tiktok accounts (left-leaning or right-leaning as set up initially).

Independent Tags	Liberal Tags	Conservative Tags
#politics, #conservative, #liberal, #news, #democrat, #republican, #USA, #politicalparties, #police, #protest, #riot, #covid, #covid19, #vaccine, #abortion, #immigration, #uspolitics, #politicians, #economics, #supremecourt, #mask, #inflation, #politicstiktok, #tax, #americanpolitics, #democracy, #senate, #climatechange, #midterms, #scotus, #government, #president, #potus, #racism, #notmypresidnet, #congress, #stock, #election2020, #politician, #usapresident, #healthcare, #rally, #housingcrisis, #military, #america, #whitehouse, #crt, #mexico, #debate, #biden, #trump, #harris, #pence, #gender, #cancleculture, #redstate, #bluestate, #feminism, #race, #unemployment, #pronouns, #christian, #election, #pandemic, #politicalcommentary, #corruption, #nato, #china, #borders, #refugee, #inflation, #neutral, #centrist, #police, #protest	#acab, #defundthepolice, #blm, #policebrutality, #shooting, #1312, #endpolicebrutality, #blacklivesmatter, #abolishthepolice, #georgefloyd, #justice, #freepalestine, #blm, #defundthepolice, #lgbtq, #blacklivesmatter, #policebrutality, #equality, #georgefloyd, #stayhome, #misinformation, #antivaxxer, #getthejab, #getvaccinated, #getboosted, #maskup, #mask, #wearamask, #staysafe, #liberal, #voteblue, #democrat, #bluewave, #aoc, #bernie, #biden, #leftist, #anarchy, #socialism, #joebiden, #vote, #feminism, #eattherich, #ketanjibrownjackson, #liberaltiktok, #berniesanders, #hilaryclinton, #communism, #climatechange, #livingwage, #medicareforall, #cancelstudentdebt, #pelosi, #democratsoftiktok, #rbg, #lgbtqrights, #alphabetmafia, #capitalism, #prochoice, #marxism, #taxtherich, #gender, #trumpism , #obama, #policereform, #genzforchange, #dontsaygay, #woke, #progressive, #leftwing, #systemicracism, #yallidarity, #classcism, #libtok, #feelthebern	#backtheblue, #bluelivesmatter, #humanizethebadge, #policeoftiktok, #thinblueline, #serveandprotect, #wedoitforyou, #policelivesmatter, #copsoftiktok, #alllivesmatter, #vaccine, #freedom, #antilockdownprotest, #freedomrally, #truckerprotest, #freedomconvoy2022, #endthemandates, #truckersoftiktok, #mandate, #masks, #mandate, #antivaxx, #stopthemandate, #medicalfreedom, #freedom, #stopthevaccine, #standupforyourrights, #myocarditis, #willnotcomply, #pureblood, #maga, #republican, #conservative, #makeamericagreatagain, #letsgobrandon, #votered, #redwave, #trump, #snowflake, #factsoverfeelings, #trump2024, #trump2020, #foxnews, #patriot, #tedcruz, #conservativetiktok, #freespeech, #republicanhypehouse, #donaldtrump, #climatehoax, #conservativeheat, #libertarianism, #libertarian, #secondadmendment, #guns, #sleepyjoe, #gop, #trumptrain, #prolife, #trumprally, #benshapiro, #saveourchildren, #proudamerican, #buildthewall, #conservativewoman, #republicangirlsdoitbest, #sheep, #todayisamerica, #wethepeople, #alm, #saveamerica, #2ndadmendment, #americafirst, #magaforever, #vaxx,

Figure 6.1: The variation across hashtags that showed up in liberal, conservative, and independent TikTok users as per the work by Lund et al. (Lund and Zhong, 2018).

2. Search the <daily keyword> as mentioned in Table 6.1.

3. Choose a video to watch from either:

    - The top 10 videos that appear on the screen after you search for the <daily keyword> or
    - Scrolling below after you click on any of the videos on that first screen (no engagement with the first video) or
    - Using the hashtags or other labels on top of the screen when you search the <daily keyword>.

4. Watch the chosen video and ask yourself: "How am I using the TikTok design features to assess its credibility and decide if I wish to share the content (or not)?"

5. Finish the engagement with the video, pause to think about what you did, whether you found it shareable, and why.

6. Now, note down your engagement procedure and your thinking process. Write it down in the provided sheet. This entry (refer to Table 6.2) should be detailed enough to clearly indicate how you employed the TikTok design affordances to assess the credibility of the content.

7. Next, note down all the metadata about the video (content description, number of likes, number of comments, number of saves, number of shares) and the TikToker (TikToker's handle, number of followers, number of following) and any further notes about the profile.

Analysis

All of the seven participant-researchers logged their engagement with TikTok videos from both accounts (politically left-leaning and politically right-leaning) at least twice a week for a total of 5 weeks between April 27 and May 28 of 2023. Collectively, the data consisted of 268 entries. These entries were unique participant-researcher impressions but might contain impressions of the same TikTok videos by unique participant-researchers.

For analyzing the 268 entries, first, we (*i.e.*, all the participant-researchers including the lead researcher) began a top-down analysis by using the findings summarized in Table **??** from Chapter 5. We used the different identified problematic behaviors (*e.g.*, "questionable purpose") as labels to categorize the 268 entries.

Day#	Date	Daily keyword
1	04/27/2023	vaccination
2	04/28/2023	climate change
3	04/29/2023	guns
4	04/30/2023	trans rights
5	05/01/2023	supreme court
6	05/02/2023	abortion
7	05/03/2023	police
8	05/04/2023	election
9	05/05/2023	immigration
10	05/06/2023	mental health

Table 6.1: Daily search keywords for the first ten days. These words were repeated in the same sequence for the remaining days for a duration of four weeks.

Daily keyword	Video Description	How did you engage with the video?
Supreme Court	A lady describes five major news updates we are waiting for from the Supreme Court on topics like student loans, affirmative action, etc.	My first impression was that this was probably an influencer-type video, but I soon realized it was very informative and journalism-like. Given that, I did not want to check its credibility and just wanted to confirm if the person was real. Upon checking the profile, it said it is from the SCOTUS blog - which I am not sure is anything official. But after scrolling through, the content primarily surfaced similar types of content on political topics and mostly informative in nature (after watching two more videos chosen at random). IRL, I would have followed this TikToker as they seem to create good and credible content, which limits to being informative and not individual opinions.
Abortion	A lady crying about how she felt sad after the abortion and how it makes you reconsider doing it,	I saw the whole video thinking if she was pro-abortion or anti. It can be twisted both ways, though I think her intent was probably promoting that abortion is hard. But then again, the caption said, "After getting 1 M like on the earlier video, I decided to share this..". I don't like that the virality of another video influenced this. The video content also had negative framings, like the career pressure will increase now, imagining how life will be with the kid, etc. I checked the profile and saw she is a cancer patient undergoing chemo (a recent development), which made me feel empathetic toward her. This kinda changed my impression of her - though she is a vlogger, she definitely cares about impressions.

Table 6.2: Table 2: Sample data records for two search terms and the video content description.

Unfortunately, we realized during the process that many labels are too specific to Twitter and do not befit the unique affordances of the TikTok platform.

For the second round, therefore, we collaboratively conducted a grounded, interpretative analysis using an affinity-diagramming exercise to recognize the common patterns across the logged impressions of the TikTok videos. During this exercise, all the participant-researchers read aloud the documented impression of engaging with a video (logged by one of them), and collectively interpreted what the unique affordance that was employed for credibility assessment of the content. We recorded a summary of the collective interpretation on a sticky note and organized them using affinity diagramming into 13 primary clusters and 64 sub-clusters spread across them. We also annotated each interpretative sticky note with a code that maps it directly to the logged impression of a TikTok video.

6.4 Findings

I now present the salient themes that stood out from across the 13 primary clusters and 64 sub-clusters spread across them that emerged during the affinity diagramming exercise. These themes will help answer the research question of identifying affordances that facilitate learning about problematic behaviors on TikTok for safe discernment of misinformation beyond Twitter, like TikTok.

6.4.1 Referencing sources outside the main video content to infer credibility

Stitching is a unique design feature offered by TikTok that offers its users the opportunity to take someone's content and add your content to reinterpret it. TikTok defines it as "a way to reinterpret and add to another user's content, building on their stories, tutorials, recipes, math lessons, and more (Newsroom, 2020)." I found that adding more relevant content through stitching — *e.g.*, diverse and multiple voices of several citizens and news, as highlighted in the example below — offered the content higher credibility and increased the subsequent shareability of that video.

> "I have heard bits of the bill in the news but was unfamiliar with it. Thus, I was inclined to believe the video but did not want to believe it blindly. I looked at the comments, but as in similar videos, all the top comments were liked, which may have prevented me from seeing

> critical content. I went to the creator's profile and saw that all the content was just coverage of Latin culture and current events. I also saw that the follower count was over 800 thousand, which made me trust them more. I went back to the video to try to figure out what the creator was trying to say. Judging by the tone and language, I inferred they were trying to spread awareness while commenting on the issue. This aligned with the content that I saw on the profile. **I also saw that the video stitched together large amounts of footage from citizens and news to add to their point.** I stopped interacting there, but I would probably share this video to talk to someone about what is happening in Florida. While the comments were unclear, the profile made the content more credible and put it into a clearer context. The many videos also added to this." - week 3: logged entry 37

Researchers believed that having a stitch or even a duet, another TikTok feature, with supporting content could have sometimes been helpful to verify the original source (week 2: logged entry 38). Stitches are often used to add to a video's original point, answer a prompt mentioned in the original video, or start a discussion about the original video's subject matter. Duets, on the other hand, are used to showcase two videos at once and often feature complementary content to the original video they Duet. Having the original duet content deleted when researchers tried to trace back to it or having that content in incomprehensible language served as a signal of distrust to the researchers (week 3: logged entry 38).

Upon combining the duet and stitching features with the related stories feature on TikTok, researchers found that contradicting details about an incident in the different videos can confuse users and create a sense of escalation when assessing content credibility. It is important to note that such a combination of referencing more content through stitching and/or duet and offering content clusters based on sound filters and topic similarity is unique to TikTok. The highlighted part in the example below shows how the researcher combined these two offerings of TikTok to judge the non-shareability of that video.

> "I scrolled four videos down on the videos page. I clicked on the comments and original sound and looked at the TikTok page. I searched on TikTok, "El Mirage Elementary School," then watched other videos from FoxNews and personal accounts on the videos page; **I became confused if the source was credible or not due to contradicting details I saw in other videos. The footage in the video also does not have a cohesive or clear timeline. Instead, it is**

> **snippets of chaotic events.** I would not share this video. I could not understand the event clearly, and the resources I found to measure the credibility by using other design affordances made me more confused. It seems to perpetuate a culture of fear, isolation, and lack of security. It makes me feel disappointed in America, but also unsurprised." - week 2: logged entry 11

Contextual stitching was another phenomenon where researchers made references to content outside the video and used it to their own advantage; in some cases, researchers found that the users borrowed upon popular trends either merely trying to use it to get more attention or trying to make made fun of it by reframing and often counter-framing the narrative. Such attempts did not explicitly use the stitching feature, as seen in the example below.

> "I knew it was a joke because of the music and the sudden cutoff at the fall. I could also tell because the video description said, 'More people need to be talking about this,' I generally associate that with being used ironically. I found the video funny, especially since I know the original video was trying to make fun of (The original video showed a woman who could not walk after getting the vaccine). However, I think even if I had not known the video it was referencing, I still would find it funny for the general critique of anti-vaxxers. I would share the video for the humor, but as it was a joke, I would stop interacting." - week 4: logged entry 3

When referencing content not in the video in consideration, researchers sometimes felt the need to go out of the TikTok platform to verify the arguments made in the video. Such attempts included going to third-party news aggregators like Google News or other news sources (week 3: logged entry 5, week 2: logged entry 25, week 3: logged entry 40, week 3: logged entry 4, week 4: logged entry 40). This practice of referencing third parties to verify the content is, however, common when it comes to verifying credibility (Khan and Idris, 2019).

6.4.2 Assessing integrity of a TikTok account

One of the important ways in which users make judgments about the integrity of an account is by inspecting if there was any consistency between the potential purpose of that account and how the content and/or the account was presented. For example, in some instances, having a consistent uniform in all the posted videos

— police uniform (week 3: logged entry 22), nurse uniform (week 3: logged entry 17) — made users perceive the accounts to have higher integrity; that wasn't always the case though. In the example below, the researcher found consistency between the purpose of the account and how they presented themselves to negatively impact their perception of that account — in particular, as the account seemed to normalize gun violence.

> "The video showed how to clear a jammed gun using a fake gun. The video's creator said his mother bought the guns for her kids, but they were too realistic, so they gave them to him. Even though I knew the guns were fake, they looked real. It made me feel anxious to see someone holding a gun. I looked at their profile, and they post videos of them using guns in a gun range. I also read the comments out of curiosity. **I would not post the video because, even though they use guns safely, I fear being harmed by a gun in the future and the negative consequences of normalizing a culture of gun ownership**." - week 3: logged entry 31

In the example below, the account that presented itself as a medical professional account published different interviews that suggested a stance on abortion that compromised the researcher's trust in the content. Similar altering of an account's credibility also occurred based on how the content and whether the generated commentary (highlighted by TikTok) supported it or opposed it. Other examples include week 2: logged entry 15, week 4: logged entry 31, and week 2: logged entry 55.

> "As this was an interview-style, there wasn't anything to really worry about credibility. I did look at the video that the comment was from, and it was of a different woman talking about her experience receiving an abortion as well. I noticed nothing in the description except that it was replying to the comment, so I was curious why this appeared in my search. When I exited the video, I noticed that **below the post on the video, it showed what she first said ("I got an abortion") and then switched to things she said later on**, and abortion was in bold. I thought it was interesting that it was going based off of the content and would not have thought that TikTok took that into account when organizing the feed." - Week 4: logged entry 27

6.4.3 Recognizing contentious behavior of a TikTok account

When assessing the credibility of the video content that researchers were observing, sometimes they found instances where the TikTokers were creating the content to maximize the views. One such example included:

> "It was a comic take that reduced my attention and engagement level with this video. I don't think I will care much about this. I check the handle, and it says Hispanics of TikTok, so that kinda adds up now. **It only has about twenty-ish videos and still managed 7.5M likes, so that makes me a bit suspicious about the account and how it has been gaining this popularity**; perhaps it was through some made-up actors sharing it too much?!" - week 2: logged entry 31

Other strategies included exclusively stitching sources that offer high popularity, *e.g.*, significant media channels, highly influential/popular TikTokers, etc. Participant-researchers found instances where there was a shock value delivery by changing the content's tone towards the end of the video or including a passive-aggressive commentary. Given that the TikTok platform obscures the point in the video that contributed to its virality, researchers found such videos do not offer much value in increasing the credibility of that particular content.

In yet another strategy, researchers also observed cases where TikTokers engaged in seemingly purposeful use of captions and/or terms that can help the creator get around specific content-ban, thus suggesting an ill tactic not to get their content banned from the platform or to ensure getting enough promotion by the platform towards being the top content. In one such example, TikToker used the phrase "toy gun" to post gun-related content that otherwise could have been banned in the context of use. It is important to note that researchers in this exercise only engaged with mostly popular videos at the top of the search results page.

> "I first wanted more context for the video, so I went to the comments. All comments discussed what they would do, so I returned to the video. **I saw the caption said "*TOY GUN*", but I thought that given the video's text, it was more of a trick to prevent the TikTok algorithm from blocking the video**. (I have seen this on TikTok, an example being people saying FAKE BODY if there was a video depicting injury). I looked at the hashtags and saw tags including #homeinvasion, #2ndamendment, and #firearmsinstructor. This made me believe that the purpose of the video was to say that it is important to have guns for protection in case

someone tries to break into your house. This matches the question posed in the video: How would you respond? The video's answer seemed to be guns. With this understanding, I returned to the comments and saw many people mentioning printing T-shirts. I realized that they were mocking how people made T-shirts of shooting victims and were used to advocate for gun control. I then realized that this meant most of the comments would have shot the person then and there without question. This was upsetting to me, especially given the lack of context. I chose not to interact with the video and the content anymore." - week 3: logged entry 8

6.4.4 Getting (in)complete context through different cues

Temporal cues

Given that the TikTok platform does not stress on time as other platforms, researchers found that this lack of temporal transparency creates a lot of chaos that users find different to understand. For example, top videos on the search page can sometimes be older with lesser engagement than a more recent video; this lets the researcher, in the case below, question how the TikTok algorithm works when deciding which content to highlight on the main search result page.

"The person was running for a student election and used glasses and a painted sign, so there was no real misleading information. **This was the third video in the top section, and I am surprised as it is from March and has low likes, shares, and comments compared to other videos further down that are focused on the US election and posted sooner** (example video from MSNBC of Joe Biden for 2024 election posted in April). I wish there were a reason why I'm seeing this button, as this is not the video I would have expected to see in the top section when doing this search. I'm not sure why it is this high and would want to know and be able to adjust so I can see the content I would actually be looking for." - week 3: logged entry 66

Having no alignment between the content date in a video and the dates in comments (and description) also made users question the credibility. This also led to instances where there were highly popular comments made by my users that seemed to be old (as compared to the video date realized upon inspection); this rendered even the good content irrelevant.

> "My first thought was that since the man was in the army, it made sense why he would be forced to get a vaccine shot. It annoyed me a little because he kept looking back at the camera with a dismissive face, as if he was complaining and thinking the whole thing was stupid. I clicked into the comments, and all the ones I could see were telling him essentially, "Deal with it" and "Stop complaining." I went to his profile and saw that the recent content was all body-building content, so it could have explained why he was so opposed to getting the shot. However, I still didn't like the video's message, so I would not share it. I also **looked at the date and saw that it was from almost a year and a half ago. This also made it even more irrelevant to me**." - week 2: logged entry 50

Audio cues

Audio is a very important part of the TikTok platform as researchers found that the tone of audio by itself can set the video to be serious or humorous or, in some cases, facilitate altering the tone of the message. Video sound also offers a new mechanism for clustering content with similar audio, providing context to the video or the more significant trend with similar video postings. Given the unique affordance of TikTok to offer such a clustering on a page not only based on content but also music, glancing at the collection of potentially similar or thematically related videos allows users to fetch any missing context and complete their assessment of that content's credibility.

Textual cues

Researchers found that TikTokers use the description, caption, and comments related to a video to infer the complete context of the video. This was in addition to using the affordances of stitching multiple videos to get extra information about the content in consideration, as seen in an earlier section. Similarly, Tiktokers can use the event-related details in the caption to complete the picture and infer the event's credibility. It was especially useful if there were several details (*e.g.*, date along with the location; multiple sources and the evidence from each in there, etc.) that complement each other. At times, the descriptions matched very well with the search bar suggestions, thus making it easy for a platform user to glance through the related content with a click.

6.5 Discussion

Given that this research started with newly created TikTok accounts seeded with specific hashtags, the TikTok algorithm didn't have much opportunity to personalize the suggested content beyond the hashtags researchers used. As a result, the findings facilitate investigating TikTok without much algorithmic interference. The high focus on trends to drive content on TikTok makes it questionable when the users come across information that could be harmful at the mercy of algorithmic mediation. This study provided an insightful opportunity to demonstrate how individuals can identify the socio-technical opacity of information as signaled by TikTok affordances, make credibility judgments, and protect themselves from believing and further sharing it. As I discuss some of these credibility judgment processes, I highlight how — even without any researcher imposition of provenance-like conceptualization — several of these processes implicitly reference information's provenance.

6.5.1 Unique affordances cater to users with higher media literacy

With the massive viewership on TikTok and the potential to misinform users, researchers and designers need to be cognizant of how TikTok supports its users to assess the credibility of information. Findings from this study indicated that the unique opportunities that TikTok offers its users to interact with the content — *e.g.*, to share or build upon any existing trends or viral videos — also affords its users distinct opportunities — *e.g.*, considering audio cue to group thematic content together — when it comes to assessing the credibility of information in the videos. I believe that one of the main reasons this has become possible is the media literacy of TikTok platform users.

Like Instagram and Snapchat, the TikTok platform has a relatively young user base, with most users being 18 to 29-year olds Auxier and Anderson (2021). This tendency to attract a younger audience and emphasize creative short-form videos gives TikTok a distinct advantage over Twitter, which has a more diverse user base and often serves as a platform for real-time updates on various content types, not limited to entertainment. The participant-researchers of this study were representative of a relatively younger population who grew up navigating online platforms and may also have had more exposure to media literacy education (Prensky, 2001).

The higher familiarity with fact-checking information and critically assessing information credibil-

ity (Helsper and Eynon, 2010; Coskun, 2021) allows TikTok users to employ unique platform features to understand why they see what they see. For example, timestamps (when displayed) of top comments helped assess why the TikTok algorithm pushed some information to the fore or if another reason was behind such a selective promotion. In another instance, researchers of this study observed that some users of the TikTok platform adopted trick strategies to avoid a content ban or get more attention to build resistance towards algorithm-promoted content. Prior research maps such behavior to a new algorithmic folk theory of social feeds called "The Identity Strainer" theory (Karizat et al., 2021). This theory describes instances when users believe an algorithm filters out and suppresses certain social identities. It describes how users change their behaviors to shape their algorithmic identities as they see fit.

6.5.2 Moving away from the unilateral focus on influencers

The primary focus of TikTok on its influencers gives immense power to the perceived popularity of a single person. In this study, I found that oftentimes, a large viewership of a video and/or comments can serve as a signal of credibility to some extent. Unfortunately, this primary attention to popular influencers has been found to give an uncritical impression that science is an individual endeavor rather than a collective enterprise (Zeng et al., 2020), contrary to the idea of critical science rooted in collaborative discovery.

The unilateral focus on an influencer post without much context behind the information in that post makes it essential for users to assess the credibility of that information. The unique affordances of TikTok offer some opportunities for users to do that. For example, we found that users can use the features of audio-based filtering or search-term-based clustering to get more context into the video topic and/or to ascertain if the video they watched offers only a small but timely and credible glimpse from a much larger event. This attempt to contextualize information suggests TikTok users' need to examine the broader network in which the information they see propagates and the composition of that information's provenance. The other features of Duet and Stitching also served as mechanisms to facilitate verifying content based on referencing the previous sharer's integrity and credibility — again asserting the importance of investigating information's provenance.

The strategies that users employed to derive a meaningful credibility signal from the available TikTok cues ascertain that at least some users of the TikTok platform are cognizant of the need to dig into the

socio-technical opacity of the information they observe on TikTok. Unfortunately, in the current platform design, the onus of employing these opportunities lay with the platform's users without any direct support to inculcate these affordances in the short time span of a user deciding to share the video. Thus, there is a clear need to support these processes and make them happen naturally as a part of TikTok media literacy so users do not need to exert themselves or struggle to make sense of the signals themselves. It is important to note that while discovering the socio-technical opacity of information, users themselves made references to information provenance without the researchers using any explicit provenance-based design provocation like in the earlier Chapter 5.

Chapter 7

Discussion: Information provenance as a cue for credibility assessment

7.1 Introduction

When answering the research question of how to (re)design online media platforms to impart in users the ability to look through the opacity of information, I introduced three studies in this dissertation on Google, Twitter, and TikTok platforms. In these studies, I answered how information provenances can vary in credibility and identified cues that can assist users in looking into the opacity and assessing information credibility meaningfully. In this chapter, I synthesize my earlier findings and address how we can generalize the cues across different platforms to signal the opacity of information and support users towards safe discernment of potentially misleading information.

The chapter is structured as follows:

1. First, I consolidate the findings from all studies and introduce a comprehensive framework. This framework elucidates the utility of information provenance across platforms for accessing contextual information and enabling users to make informed judgments about information credibility.

2. Second, I delve into how information provenance can more generally serve as a construct to assist users in understanding the opacity of information across multiple platforms.

3. Third, I set the foundation for incorporating provenance into the design of media platforms. I present specific design opportunities aimed at assisting users in exploring the socio-technical context of information as a fundamental aspect of their credibility assessment process.

4. Fourth, I emphasize how collaborating with technology platforms to integrate information provenance can promote media literacy. Such a collaboration is crucial for supporting user efforts to assess information credibility and discern potentially misleading content effectively.

7.2 Assessing credibility across Google Search, Twitter and TikTok platforms

In Chapter 4, I demonstrated that different provenances of information that surface within a platform (Google Search) can differ in their socio-technical context — *e.g.*, different media channels and/or medium of information — and thereby information with varying credibility. Next, in Chapter 5, I demonstrated that design conceptualizations that make information provenance salient on a platform (Twitter) can help users interpret the socio-technical context of that information. Finally, in Chapter 6, I identify how the socio-technical context of information provenance can be signaled using the available affordances of a distinct platform (TikTok). Despite the differences in these platforms' socio-technical operations, the common thread across these findings is the role of information provenance in providing contextual information, which users can use to evaluate content credibility. I will now delve into how information provenance can facilitate credibility assessment on different platforms, specifically Twitter and TikTok, and then present a unified framework encompassing these platforms.

One of the initial methods users employed to gauge the credibility of information was by assessing the authenticity of the account presenting it. The perceived legitimacy of an account often hinged on whether its stated purpose aligned with the type of content it shared. Although I did not conduct user-based assessments on the Google Search platform, there were instances where advertisements, disguised as coming from genuine accounts, later turned out to disseminate delegitimizing information. On Twitter, understanding an account's purpose helped participants anticipate the type of content it might produce or promote. Any inconsistency between the perceived purpose of an account and its posted content could raise suspicions of

inauthenticity, prompting users to scrutinize both the account and its content more closely. Similarly, users on TikTok focused on assessing consistency, but they paid greater attention to the format and presentation style, given the platform's emphasis on video-based content.

Furthermore, users on both Twitter and TikTok shared a common focus on identifying contentious behavior. On Twitter, users evaluated this based on the distribution of an account's activity, such as the ratio of tweets to replies, as indicated in cue cards. This approach helped them discern any toxic behaviors associated with an account. Participants also recognized that retweets from recently created Twitter accounts could contribute to the account's contentious behavior. This concern, combined with any suspicion of the account's authenticity, hinted at potential involvement in organized efforts to incite discord. Conversely, when inferring contentious behavior on TikTok, users referenced more nuanced signals, such as attempts to stitch together videos from popular TikTokers to maximize views and avoid content restrictions by using deceptive terms in video captions.

The most important way in which these platforms differed when it came to assessing content credibility was how they afforded background context of the presented information. On Twitter, users leveraged the platform's network-driven nature to evaluate the online connections of an account when assessing information from it. By examining the content created by the account, along with the accounts it retweeted and the hashtags and URLs it shared, users could gauge the online associations of that account. This helped users understand the community the account belonged to and the type of content it might endorse and amplify. Users could then employ these judgments to evaluate the information shared by the actors in the context of their online network (Fleming, 2014). I believe that search engine platforms might also support their users towards credibility assessment by offering more salience to some of the contexts behind information provenances as seen by those users on the search engine home page.

The process of discrediting information from an account based on incomplete or suspicious context differed significantly from the process described above for Twitter. It was primarily driven by users' desire to triangulate various pieces of information as they assessed credibility. This involved noting the timestamp of the information (or its absence) to establish a clear timeline of events, listening to embedded audio to contextualize trends, and matching captions with search terms to explore similar events and activities discussed on the platform. These processes leverage the unique affordances offered by TikTok, providing

Table 7.1: Users employ different the notion of provenance on different platforms to look into information opacity and make credibility judgments along several dimensions.

Dimension of credibility assessment	**Twitter Affordances**	**TikTok Affordances**
Authenticity of an account	Questionable purpose	Questionable purpose
	Inconsistent participation	Inconsistency between format of videos and topical alignment
Inferring complete context	Lack of shared interests	Triangulation across clustered content
	Unfavorable connections	Setting up a clear timeline through timestamps
	Uncredible media sources	Match between search term-based suggested videos
Contentious behavior of an account	High #retweets/#replies	Low #videos but high views
	Getting retweeted by new accounts	Stiching only popular TikTokers
		Using ban-evading tricks
Trust towards info-spreaders	Unfamiliarity in one's network	Stitching videos of TikTokers with low credibility
	Lack of expert popularizers	Contextual stitching to alter topicality/tone
Consequences of sharing	Potentially create problematic association	

users with new cues to evaluate content credibility.

Another point of differentiation between these platforms was in how they assessed distributed trust in information spreaders. On Twitter, this process was relatively straightforward: users checked who shared the information as it gained popularity. On TikTok, however, this process was more complex, as users combined two unique features: duet and stitching. By combining these features with the related stories feature on TikTok, users discovered that contradictory details about an incident in different videos could confuse users and contribute to a sense of escalation. While Google Search currently lacks a mechanism to identify the agency behind popularized information on its page, it remains to be seen how this might change as influencer-driven content becomes more prevalent. Table 7.1 provides a summary of how these findings converge for Twitter and TikTok platforms.

7.3 Information provenance: A platform agnostic construct to convey information opacity and assess credibility

User actions and algorithmic decisions play a pivotal role in governing information dissemination across platforms. This dissertation explores a range of media platforms that utilize both user and algorithmic agencies to varying extents. While Google's search engine remains largely unaffected by individual search behaviors, Twitter (as of 2020) relies heavily on actions taken by followers to shape information feeds. TikTok occupies an intermediary position, blending user actions with virality-driven decisions to recommend relevant content. Despite the diversity in information curation approaches, my research underscores the significance of leveraging the concept of information provenance to gain contextual insights into information dissemination. Obtaining such insights is crucial for reducing the perceived opacity surrounding how information spreads across these platforms.

Previous research across various disciplines, such as artificial intelligence (Hase and Bansal, 2020; Shin, 2021; Eslamimehdiabadi, 2019), privacy (Hargittai, 2007), and cryptography (Peeters and Pulls, 2015), has consistently shown that providing transparency into system functionalities and decision-making processes can enhance users' perceptions of system credibility. In this dissertation, I build upon this scholarship by demonstrating how information provenance effectively aids users in understanding the dissemination of information narratives within online environments, thereby reducing perceived opacity.

My research presents provenance as a tool for helping users meaningfully engage with complex information ecosystems, aligning with the perspectives of Vallor (Vallor, 2016) and Badke (Badke, 2021). While the core objective remains consistent—understanding how specific information reaches users' feeds—the focus varies across platforms. On Twitter, opacity relates more to who in the user's network amplifies information, while on TikTok, opacity refers to the challenge of understanding why the algorithm prioritizes certain videos [1]; the same opacity on TikTok referred to the inability to reason why the algorithm is pushing a certain video to them.

The realization that information provenance is key to understanding the propagation of information is agnostic to the choice of media platforms, as seen in this research through the diverse interpretation of

[1] Following Twitter's acquisition, Elon Musk introduced a "For You" feed, granting algorithms greater influence over content recommendation beyond traditional follower-following dynamics (Perrigo, 2023)

provenance. Recent research underscores this indifference across platforms; *e.g.*, researchers have adopted the construct of provenance to offer transparency into the sequence of visual edits made to a media file when assessing the relationship between credibility and user exposure to the sequence of those edits (Feng et al., 2023). Irrespective of the platform, information provenance acts as a key indicator that diminishes opacity, offering a valuable mechanism for assessing information credibility.

7.4 Designing platforms centered around information provenance

To tackle the complex challenge of assessing credibility in the face of potentially misleading information (Jack, 2019; Wardle, 2018; Rittel and Webber, 1973), we require solutions that grasp and convey the socio-technical elements underlying it. Therefore, it is critical to incorporate the construct of information provenance into the design of online media platforms (within reasonable constraints) to foster a collaborative partnership between online media platforms and their users, as I discuss below.

7.4.1 Facilitate easy access to online information provenance

While existing approaches to assist users in investigating information propagation have shown effectiveness (Finn et al., 2015), it's imperative for platforms to integrate them in user-friendly ways that align with everyday scenarios, such as retweeting on Twitter or forwarding on Whatsapp, where quick decision-making is necessary. In my research, one approach I explored involved an intervention comprising only three accounts: the root, a popularizer, and a friend, as detailed in Chapter 5. This trajectory design prompted participants to evaluate their confidence in the propagating accounts and the potential consequences of sharing the information further. I believe that by implementing different design choices, such as conceptualizing provenance as an 'information accordion' that offers on-demand access to more propagation factors, further unique insights facilitated by information provenances can be discovered.

However, there is concern that providing access to specific accounts and their activity could lead to partisan sorting and increase online polarization (Törnberg, 2022). To address this, it may be beneficial to include instances where ideologically similar or familiar users share diverse or dissenting perspectives within the provenance. For example, incorporating popularizers who use the quote-tweeting mechanism on Twitter or influencers who share posts with contrasting captions on Instagram could provide interesting but

contradictory perspectives. Participants from studies detailed in Chapter 5 and Chapter 6 indicated that such accommodations within the design of provenance could encourage the consideration of alternate viewpoints.

7.4.2 Provide cues that signal account's evolving behavior

Online media platforms require cues that accurately reflect the genuine activity of user accounts, preventing them from easily misrepresenting themselves (*e.g.*, in their profile descriptions on Twitter and TikTok) without drawing scrutiny from other users. Information provenances offer users insights into various concerns related to shared information, as outlined in Table 7.1, prompting them to question the trustworthiness of content shared by certain accounts. By incorporating provenance-based cues and more nuanced variations, users can reflect on their online sharing practices and address problematic accounts effectively.

While the research presented in this dissertation primarily focuses on the sharing of information, the conceptualization of information provenance has broader implications. It can assist practitioners and researchers in developing contextual cues for a range of decision-making processes based on an account's background, history, and context of activity. For instance, users may consider these cues when deciding to follow/unfollow or friend/unfriend an account, or when deleting past engagements with an account. As platforms continue to evolve, such a framework can facilitate the design of tailored contextual cues that help platforms expose and counter sophisticated mechanisms for promoting problematic content.

7.4.3 Provide cues that vary as per the actor's role in information provenance

I found that participants asked different questions when reflecting upon the different accounts involved in information propagation. 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 unfamiliar account — only when that account plays a major role in spreading information. For within-network platforms, 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 in promoting 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 (Hancock, 2020; McCrosky, 2020), 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 2020 took upon the initiative to verify the accounts of several medical professionals to identify accurate COVID-related information (Gadde and Derella, 2020).

7.4.5 Communicate information spread to users of variable information literacy

The information spread on social media is known to be networked. Though more advanced users of these media platforms, as seen in the case of emergent platforms like TikTok and Snapchat, might understand and are well accustomed to dealing with such networkedness and its potential implications, more average users of these platforms could struggle to infer even the most basic lessons by themselves. This limitation mandates platforms to devise ways in which they can communicate the socio-technical opacity of information to users that suit their 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. Enhancing the visibility of information provenance stands out as a potential strategy to highlight the opacity of information and facilitate credibility assessment.

7.5 Rethinking media literacy by incorporating information provenance as a platform feature

This research began with an understanding of information provenance as a documented record tracking the origin and movement of information across users and platforms. Throughout this process, I observed that as researchers and users engaged with various platform cues to assess information credibility, the concept of provenance expanded beyond mere information transfer. For instance, on Google Search, search character-

istics provided rich context to the provenance of searched information, impacting its credibility, particularly in relation to election headlines. Similarly, on Twitter, the prominence of an influencer's account within the provenance affected how users evaluated information credibility. Even without deliberate design interventions, TikTok users extended the notion of provenance to include modifications driven by captions and voice-overs.

These observations led me to believe that a focus solely on the passage of information between users cannot define information provenance comprehensively. To capture the nuanced socio-technical perspective of provenance as seen through these studies, it becomes critical to incorporate the processes of creating (*authors!*), processing (*algorithms!*), modifying (*edits!*), observing (*access!*), interpreting (*analysis!*), and transferring (*shares!*) information as a part of its provenance — processes which very closely align with the definition of media literacy!

While the studies in this dissertation focused on specific platforms, their findings and implications extend to various online media platforms. Information operations observed across these platforms are highly interconnected (Röchert et al., 2021; Wilson and Starbird, 2020). However, techniques to combat misinformation, such as credibility assessment, content moderation, and fact-checking, do not uniformly benefit from this interconnectedness. Moreover, false stories tend to spread at least ten times faster than true ones (Vosoughi et al., 2018). Accessing provenance can help users understand the opacity of information propagation, enabling them to evaluate information based on content and the larger ecosystem promoting it. In crucial times when these ecosystems rapidly evolve through organized efforts (Wilson, 2021), the transparency provided by provenances into the spread of information can help users make credibility judgments and consider the implications of sharing content, potentially curbing the spread of misinformation.

Thus, it is possible to achieve a true sense of media literacy through a partnered effort between platforms *to exercise provenance-based design features to support users for looking into informational opacity* and users *to get familiar and responsibly engage with those design features*. However, just as there are challenges to steering platform design choices, getting users accustomed to the new features is tricky. Besides, getting users adept at using the new features may not necessarily solve the challenge of effective credibility assessment. There is already some debate about the benefits of learning media literacy in modern times and how it might potentially backfire. An imminent scholar, Boyd, argued that while achieving media literacy is

imagined to be empowering, it is a form of critical thinking that tells people to question what they see (Boyd, 2017). This could be concerning given that incoherent critical thinking can lead to the emergence of narratives questioning the legitimacy of facts (Lantian et al., 2021). In countering the concern, critics argued that critical thinking is not necessarily rooted in positivist traditions as Boyd's interpretation suggests (Noula, 2018). In other words, while there is a danger in promoting user adoption of provenance as an essential skill to think critically, there is a benefit in promoting user adoption of provenance as a value to be reflective of one's informational environment. Cognizance of the environment in which information spreads can help users recognize the limitations of isolated media experiences and encourage them to advocate for increased media literacy efforts at the platform level (Adams and Hamm, 2001).

Chapter 8

Conclusion

Having outlined the significance of information provenance as a crucial cue for reducing information opacity, I now reflect on the limitations and challenges encountered in this research. Finally, I conclude my dissertation by summarizing its key contributions.

8.0.1 Limitations

One significant challenge in implementing cues resembling information provenance is that platforms often prioritize enhancing user engagement and information sharing over evaluating credibility. Moreover, many users prefer solutions that emphasize simplicity and a user-friendly experience. To effectively address the complexity of misinformation and its wickedness (Jack, 2019; Wardle, 2018), fostering a collaborative partnership between online media platforms and their users is essential. However, the current technological landscape, marked by concerns about regulating free speech and workforce reductions, presents obstacles to such partnerships. Without enthusiastic leadership from platforms, offering these solutions to users becomes impossible. Moreover, these solutions may fail if they do not align with the existing state of media literacy among users.

I believe that users with higher media literacy stand to benefit more from the proposed research compared to those with lower media literacy. While reflective cues can effectively promote critical thinking and reveal hidden aspects of online information, automatically derived heuristic cues can reduce cognitive load on users during decision-making (Caraban et al., 2019). To leverage this insight effectively, future designs

of provenance-based cues should prioritize ease of use. For instance, instead of displaying the entire distribution of tweets, retweets, quote tweets, and comments, a simplified icon could convey a relatively higher proportion of retweets.

Social media content and design evolve constantly to accommodate the changing trends, policy requirements, and discussions within the community. Therefore, research on online media platforms is sensitive to the time of conduct and could impose limitations on the generalizability of the research and its implications. It is also possible that a different dataset or participant sample might impact some of the findings. Despite this possibility, I feel confident that similar research activities will underscore the utility of information provenance-based solutions to understand the opacity of information.

8.0.2 Summary of contributions

Collectively, through studies on three unique sites of research, my dissertation makes several data, empirical, artifact, theoretical, and methodical contributions. One of these contributions includes a platform-agnostic understanding of how design features of different platforms can afford credibility assessment along the dimensions of inauthenticity, unfavorable online associations, contentious behavior, lack of trust, and unwanted consequences of sharing information. Most importantly, this dissertation contributes to and establishes information provenance as a platform-agnostic construct that conveys the opacity of information to users in a way that is useful for assessing information credibility.

When answering the first research question on how contextual factors of informational search on Google Search impact the credibility of information, I discovered how some information provenances on the search page could contribute more delegitimizing content than others. Although the study itself was not originally about provenance, the results were suggestive of information having provenance that was not made salient by the Google Search platform. I assert that knowing the context behind information search will let a user witness the broader context of search results and identify how different provenances may serve information that varies in terms of credibility. I also contribute a dataset of election-related headlines, a codebook to characterize the credibility of those headlines along several dimensions, and policy suggestions to aid audits of election-based information online.

I leveraged the Twitter platform to answer the second research question about how the context behind

information, made explicit using a provenance-based intervention, helps users assess the credibility of that information. I demonstrate that provenance-based cues are useful in network-based platforms to facilitate credibility assessment of information at a broader level of the overall propagation network. Supplementary contributions encompass the 'tweet trajectory' intervention and its variations, providing potential avenues for future researchers to explore propagation networks in distinctive ways.

The final study answered the third research question about identifying ways in which platforms can signal the socio-technical context of information. Based on the credibility assessment of information on the TikTok platform, this study revealed that as users employ nuanced strategies unique to platform features to afford the informational context, they make implicit references to provenance. The study also makes a methodical contribution in using a diary study with researchers as the participants to record the role of platform features to facilitate credibility assessment.

The existing design of online media platforms focuses on creating online feeds with information afforded by users' behavioral history or their online network. Unfortunately, reducing the salience of the context that channeled the information to users compromises the significance of those contextual factors for assessing credibility. I believe that incorporating cues based on information provenance will bring attention to the contextual surroundings of information, which are currently obscured by the design of media platforms. Such cues could offer users an opportunity to re-evaluate their trust in the information and, by extension, in the media platforms.

Bibliography

Crystal Abidin. 2016. Visibility labour: Engaging with influencers' fashion brands and #OOTD advertorial campaigns on Instagram. *Media International Australia*, 161(1):86–100.

Ad Fontes Media, Inc. 2020. The media bias chart 6.0. *Ad Fontes Media*.

Dennis Adams and Mary Hamm. 2001. *Literacy in a multimedia age*. Christopher-Gordon Publications.

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*, pages 431–442.

Jonathan Albright. 2018. Untrue-tube: Monetizing misery and disinformation. *Medium*.

Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

R Michael Alvarez, Jian Cao, and Yimeng Li. 2021. Voting experiences, perceptions of fraud, and voter confidence. *Social Science Quarterly*, 102(4):1225–1238.

Janna Anderson and Lee Rainie. 2017. The future of truth and misinformation online. *Pew Research Center*.

Katie Elson Anderson. 2020. Getting acquainted with social networks and apps: It is time to talk about TikTok. *Library Hi Tech News*, 37(4):7–12.

Blake C Andrew. 2007. Media-generated shortcuts: Do newspaper headlines present another roadblock for low-information rationality? *Harvard International Journal of Press/Politics*, 12(2):24–43.

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):1–27.

Dan Arnaudo. 2017. Computational propaganda in Brazil: Social bots during elections. *Computational Propaganda Research Project*.

Brooke Auxier and Monica Anderson. 2021. Social media use in 2021. *Pew Research Center*, 1:1–4.

Stefan Bächtold. 2023. Blackouts, whitelists, and 'terrorist others': The role of socio-technical imaginaries in myanmar. *Journal of Intervention and Statebuilding*, pages 1–21.

Franziska Badenschier and Holger Wormer. 2011. Issue selection in science journalism: Towards a special theory of news values for science news? In *The sciences' media connection–public communication and its repercussions*, pages 59–85.

William Badke. 2021. *Research strategies: Finding your way through the information fog*. Iuniverse.

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):9216–9221.

Andrea Ballatore, Mark Graham, and Shilad Sen. 2017. Digital hegemonies: The localness of search engine results. *Annals of the American Association of Geographers*, 107(5):1194–1215.

Andrew J Baranauskas. 2022. News media and public attitudes toward the protests of 2020: An examination of the mediating role of perceived protester violence. *Criminology & Public Policy*, 21(1):107–123.

Pablo Barberá. 2020. Social media, echo chambers, and political polarization. *Social media and democracy: The state of the field, prospects for reform*, 34:34–55.

Pablo Barberá and Gonzalo Rivero. 2015. Understanding the political representativeness of Twitter users. *Social Science Computer Review*, 33(6):712–729.

Corey H Basch, Zoe Meleo-Erwin, Joseph Fera, Christie Jaime, and Charles E Basch. 2021. A global pandemic in the time of viral memes: Covid-19 vaccine misinformation and disinformation on TikTok. *Human vaccines & immunotherapeutics*, 17(8):2373–2377.

Nicholas M Baumel, John K Spatharakis, Steven T Karitsiotis, and Evangelos I Sellas. 2021. Dissemination of mask effectiveness misinformation using TikTok as a medium. *Journal of Adolescent Health*, 68(5):1021–1022.

Halina Bause. 2021. Political social media influencers as opinion leaders? *Publizistik*, 66:295–316.

Pavel Sidorenko Bautista, Nadia Alonso-López, and Fábio Giacomelli. 2021. Fact-checking in TikTok. Communication and narrative forms to combat misinformation. *Revista Latina de Comunicación Social*, (79):87–112.

David Bawden. 2001. Information and digital literacies: a review of concepts. *Journal of documentation*, 57(2):218–259.

David Bawden et al. 2008. Origins and concepts of digital literacy. *Digital literacies: Concepts, policies and practices*, 30(2008):17–32.

Michael A. Beam. 2014. Automating the news: How personalized news recommender system design choices impact news reception. *Communication Research*, 41(8):1019–1041.

Puneet Bhargava, Katie MacDonald, Christie Newton, Hause Lin, and Gordon Pennycook. 2023. How effective are TikTok misinformation debunking videos? *Harvard Kennedy School Misinformation Review*.

Sima Bhowmik, Md Main Uddin Rony, Md Mahfuzul Haque, Kristen Alley Swain, and Naeemul Hassan. 2019. Examining the role of clickbait headlines to engage readers with reliable health-related information. *arXiv preprint arXiv:1911.11214*.

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):1–30.

Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26.

Namle Board. 2007. Core principles of media literacy education in the United States. *Published online*.

Leticia Bode and Kajsa E Dalrymple. 2016. Politics in 140 characters or less: Campaign communication, network interaction, and political participation on Twitter. *Journal of Political Marketing*, 15(4):311–332.

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):212–232.

Prashant Bordia and Nicholas DiFonzo. 2017. Psychological motivations in rumor spread. *Rumor Mills: The Social Impact of Rumor and Legend*, pages 87–102.

Danah Boyd. 2017. Did media literacy backfire? *Journal of Applied Youth Studies*, 1(4):83–89.

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*, pages 1–10.

Samantha Bradshaw. 2019. Disinformation optimised: gaming search engine algorithms to amplify junk news. *Internet policy review*, 8(4):1–24.

Samantha Bradshaw, David Thiel, Carly Miller, and Renee DiResta. 2020. Election delegitimization: Coming to you live. *Election Integrity Partnership*.

Brennan Center for Justice. 2020. It's official: The election was secure. *Published Online*.

Heather Z Brooks and Mason A Porter. 2020. A model for the influence of media on the ideology of content in online social networks. *Physical Review Research*, 2(2):023041.

Nico Brooks. 2004. The Atlas rank report: How search engine rank impacts traffic. *Insights, Atlas Institute Digital Marketing*.

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

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*, pages 1–15.

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

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)*, pages 214–219.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 675–684.

Rong-Ching Chang, Chun-Ming Lai, Kai-Lai Chang, and Chu-Hsing Lin. 2021. Dataset of propaganda techniques of the state-sponsored information operation of the people's republic of china. *ArXiv*, abs/2106.07544.

Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

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):25–35.

Sedona Chinn, Dan Hiaeshutter-Rice, and Kaiping Chen. 2023. How science influencers polarize supportive and skeptical communities around politicized science: A cross-platform and over-time comparison. *Political Communication*, pages 1–22.

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):315–333.

Harris Cohen. 2021. Helpful search tools for evaluating information online. *Google Blog*.

Julie Coiro. 2014. Teaching adolescents how to evaluate the quality of online information. *Edutopia*.

Keith Coleman. 2021. Introducing Birdwatch, a community-based approach to misinformation. *Twitter Blog*.

Andrea Cornwall and Rachel Jewkes. 1995. What is participatory research? *Social science & medicine*, 41(12):1667–1676.

Cicek Coskun. 2021. Digital literacy in the world of digital natives. In *Handbook of research on new media applications in public relations and advertising*, pages 486–504. IGI Global.

Kate Crawford and Ryan Calo. 2016. There is a blind spot in AI research. *Nature*, 538(7625):311–313.

Cybersecurity & infrastructure security agency. 2021. Election security rumor vs. reality. *Published online*.

Cybersecurity & Infrastructure Security Agency. 2022. Election infrastructure security. *Published online*.

Aaron Delwiche and Mary Margaret Herring. 2020. Propaganda critic, Russian disinformation, and media literacy: A case study. In *Media Literacy in a Disruptive Media Environment*, pages 94–108.

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*, pages 2451–2460.

Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. 2018. I vote for—how search informs our choice of candidate. *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple, M. Moore and D. Tambini (Eds.)*, 22.

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).

Gregory J Digirolamo and Douglas L Hintzman. 1997. First impressions are lasting impressions: A primacy effect in memory for repetitions. *Psychonomic Bulletin & Review*, 4(1):121–124.

Renee DiResta. 2018. Computational propaganda: If you make it trend, you make it true. *The Yale Review*, 106(4):12–29.

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. *United States Senate Documents*.

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

Gabriele Donzelli, Giacomo Palomba, Ileana Federigi, Francesco Aquino, Lorenzo Cioni, Marco Verani, Annalaura Carducci, and Pierluigi Lopalco. 2018. Misinformation on vaccination: A quantitative analysis of YouTube videos. *Human vaccines & immunotherapeutics*, 14(7):1654–1659.

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*, pages 266–285.

Ullrich KH Ecker, Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. 2014. The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied*, 20(4):323.

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):52–71.

Rosenberg Eli. 2019. How anonymous tweets helped ignite a national controversy over maga-hat teens. *The Hill*.

Elmer Emig. 1928. The connotation of newspaper headlines. *Journalism Quarterly*, 4(4):53–59.

Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521.

Motahhare Eslamimehdiabadi. 2019. *Participating and designing around algorithmic socio-technical systems*. Ph.D. thesis, University of Illinois.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* power 3: A flexible statistical

power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191.

KJ Kevin Feng, Nick Ritchie, Pia Blumenthal, Andy Parsons, and Amy X Zhang. 2023. Examining the impact of provenance-enabled media on trust and accuracy perceptions. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–42.

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

Miriam Fernández, Alejandro Bellogín, and Iván Cantador. 2021. Analysing the effect of recommendation algorithms on the amplification of misinformation. *arXiv preprint arXiv:2103.14748*.

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

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*, pages 1–2.

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):146–165.

Peter L Forberg. 2022. From the fringe to the fore: An algorithmic ethnography of the far-right conspiracy theory group QAnon. *Journal of Contemporary Ethnography*, 51(3):291–317.

Jeffrey Friedman. 2019. *Power without knowledge: A critique of technocracy*. Oxford University Press.

Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on Twitter? In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, pages 179–192.

Vijaya Gadde and Matt Derella. 2020. An update on our continuity strategy during Covid-19. *Twitter Blog*.

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

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*, pages 1–14.

Hyungjin Gill. 2022. Testing the effect of cross-cutting exposure to cable tv news on affective polarization: Evidence from the 2020 US presidential election. *Journal of Broadcasting & Electronic Media*, 66(2):320–339.

James G Gimpel, Nathan Lovin, Bryant Moy, and Andrew Reeves. 2020. The urban-rural gulf in american political behavior. *Political behavior*, 42(4):1343–1368.

Nathaniel Gleicher. 2018. Coordinated inauthentic behavior explained. *Facebook Blog*.

Google Search. 2020. See what was trending in 2020 - United States. *Google report on trending topics*.

Sonja Grabner-Kräuter and Sofie Bitter. 2015. Trust in online social networks: A multifaceted perspective. In *Forum for Social Economics*, volume 44, pages 48–68.

Justin Grandinetti and Jeffrey Bruinsma. 2023. The affective algorithms of conspiracy TikTok. *Journal of Broadcasting & Electronic Media*, 67(3):274–293.

Stephan Grimmelikhuijsen, Gregory Porumbescu, Boram Hong, and Tobin Im. 2013. The effect of transparency on trust in government: A cross-national comparative experiment. *Public Administration Review*, 73(4):575–586.

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*, pages 228–243.

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

M Nick Hajli, Julian Sims, Mauricio Featherman, and Peter ED Love. 2015. Credibility of information in online communities. *Journal of Strategic Marketing*, 23(3):238–253.

Hale Spencer, Saranac. 2020. Nine election fraud claims, none credible. *FactCheck.org*.

Jeff Hancock. 2020. Et speaker series: Rethinking trust and well-being in this strange new world. *Mozilla, Youtube*.

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):3–28.

Eszter Hargittai. 2007. Whose space? differences among users and non-users of social network sites. *Journal of Computer-Mediated Communication*, 13(1):276–297.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*.

Alison Hearn and Stephanie Schoenhoff. 2015. From celebrity to influencer: Tracing the diffusion of celebrity value across the data stream. *A Companion to Celebrity*, pages 194–212.

Ellen Johanna Helsper and Rebecca Eynon. 2010. Digital natives: Where is the evidence? *British Educational Research Journal*, 36(3):503–520.

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

John Herrman. 2019. How TikTok is rewriting the world. *The New York Times*, 10:412586765–1586369711.

Sally Rao Hill, Indrit Troshani, and Dezri Chandrasekar. 2017. Signalling effects of vlogger popularity on online consumers. *Journal of Computer Information Systems*, pages 76–84.

Brian Hilligoss and Soo Young Rieh. 2008. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management*, 44(4):1467–1484.

Renee Hobbs. 2021. *Media literacy in action: Questioning the media*. Rowman & Littlefield Publishers.

Richard Hornik and Masato Kajimoto. 2014. 'De-Americanizing' news literacy: Using local media examples to teach critical thinking to students in different socio-cultural environments. *Asia Pacific Media Educator*, 24(2):175–185.

Carl Iver Hovland, Irving Lester Janis, and Harold H Kelley. 1953. *Communication and persuasion*. Yale University Press.

Desheng Hu, Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2019. Auditing the partisanship of Google search snippets. In *The World Wide Web Conference*, pages 693–704.

Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27.

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119.

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*, pages 1–12.

The Aspen Institute. 2021. Commission on information disorder: Final report. *The Aspen Institute*.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the United States. *Annual review of political science*, 22:129–146.

Bahruz Jabiyev, Sinan Pehlivanoglu, Kaan Onarlioglu, and Engin Kirda. 2021. FADE: Detecting fake news articles on the web. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, pages 1–10.

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

Caroline Jack. 2019. Wicked content. *Communication, Culture & Critique*, 12(4):435–454.

Justin Jagosh, Ann C Macaulay, Pierre Pluye, JON Salsberg, Paula L Bush, JIM Henderson, Erin Sirett, Geoff Wong, Margaret Cargo, Carol P Herbert, et al. 2012. Uncovering the benefits of participatory research: Implications of a realist review for health research and practice. *The Milbank Quarterly*, 90(2):311–346.

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):1–42.

Nuurrianti Jalli. 2021. 'Mission impossible?': Tracking political misinformation and disinformation on TikTok. *The Conversation*.

Kathleen Hall Jamieson, Bruce W Hardy, and Daniel Romer. 2007. The effectiveness of the press in serving the needs of American democracy.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7–38.

Prerna Juneja and Tanushree Mitra. 2021. Auditing e-commerce platforms for algorithmically curated vaccine misinformation. In *Proceedings of the 2021 chi conference on human factors in computing systems*, pages 1–27.

Masato Kajimoto. 2016. Developing news literacy curricula in the age of social media in Hong Kong, Vietnam and Myanmar. *Journalism Education*, 5(1):136–155.

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

Jean-Noël Kapferer. 1987. *Rumeurs: le plus vieux média du monde*. Editions du seuil.

Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic folk theories and identity: How TikTok users co-produce knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–44.

Anna Kawakami, Khonzodakhon Umarova, and Eni Mustafaraj. 2020. The media coverage of the 2020 US presidential election candidates through the lens of Google's top stories. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 868–877.

Douglas Kellner and Jeff Share. 2005. Media literacy in the US. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, 11:1–21.

M Laeeq Khan and Ika Karlina Idris. 2019. Recognise misinformation and verify before sharing: A reasoned action and information literacy perspective. *Behaviour & Information Technology*, 38(12):1194–1212.

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

Christopher Kliewer. 2008. Joining the literacy flow: Fostering symbol and written language learning in young children with significant developmental disabilities through the four currents of literacy. *Research and Practice for Persons with Severe Disabilities*, 33(3):103–121.

Christopher Kliewer, Linda May Fitzgerald, Jodi Meyer-Mork, Patresa Hartman, Pat English-Sand, and Donna Raschke. 2004. Citizenship for all in the literate community: An ethnography of young children with significant disabilities in inclusive early childhood settings. *Harvard Educational Review*, 74(4):373–403.

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

Charles H Knoblauch. 1990. Literacy and the politics of education. *The Right to Literacy*, pages 74–80.

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.

Tibor Koltay. 2011. The media and the literacies: Media literacy, information literacy, digital literacy. *Media, Culture & Society*, 33(2):211–221.

Ava Korbani and Jessica LaBrie. 2021. Toxic TikTok trends. *Journal of Student Research*, 10(2).

Daniel Kreiss and Shannon C McGregor. 2019. The "arbiters of what our voters see": Facebook and Google's struggle with policy, process, and enforcement around political advertising. *Political Communication*, 36(4):499–522.

Elana R Kriegel, Bojan Lazarevic, Christian E Athanasian, and Ruth L Milanaik. 2021. TikTok, tide pods and tiger king: health implications of trends taking over pediatric populations. *Current Opinion in Pediatrics*, 33(1):170–177.

K Hazel Kwon, Mi Hyun Lee, Sang Pil Han, and Sungho Park. 2022. Fake thumbs in play: A large-scale exploration of false amplification and false diminution in online news comment spaces. *New Media & Society*, page 14614448221099170.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Anthony Lantian, Virginie Bagneux, Sylvain Delouvée, and Nicolas Gauvrit. 2021. Maybe a free thinker but not a critical one: High conspiracy belief is associated with low critical thinking ability. *Applied Cognitive Psychology*, 35(3):674–684.

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*, pages 1–15.

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

Junhao Li, Miikka Kuutila, Eetu Huusko, Nimantha Kariyakarawana, Marko Savic, Nazanin Nakhaie Ahooie, Simo Hosio, and Mika Mäntylä. 2023. Assessing credibility factors of short-form social media posts: A crowdsourced online experiment. In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI chapter*, pages 1–14.

Hai Liang, Isaac Chun-Hai Fung, Zion Tsz Ho Tse, Jingjing Yin, Chung-Hong Chan, Laura E Pechta, Belinda J Smith, Rossmary D Marquez-Lameda, Martin I Meltzer, Keri M Lubell, et al. 2019. How did Ebola information spread on Twitter: broadcasting or viral spreading? *BMC Public Health*, 19(1):1–11.

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

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):40–52.

Ioana Literat, Abubakr Abdelbagi, Nicola YL Law, Marcus YY Cheung, and Rongwei Tang. 2021. Research note: Likes, sarcasm and politics: Youth responses to a platform-initiated media literacy campaign on social media. *Harvard Kennedy School Misinformation Review*.

Ioana Literat, Yoo Kyung Chang, and Shu-Yi Hsu. 2020. Gamifying fake news: Engaging youth in the participatory design of news literacy games. *Convergence*, 26(3):503–516.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514.

Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, 59(7):1041–1052.

Campbell Lund and Shirui Zhong. 2018. The impact of TikTok's engagement algorithm on political polarization. *Published Online*.

Noam Lupu, Mariana V Ramírez Bustamante, and Elizabeth J Zechmeister. 2020. Social media disruption: Messaging mistrust in latin america. *Journal of Democracy*, 31(3):160–171.

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)*, pages 1–4.

Diego A Martin and Jacob N Shapiro. 2019. Trends in online foreign influence efforts. *Princeton University Princeton, NJ*.

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

Michele Mazza, Guglielmo Cola, and Maurizio Tesconi. 2022. Ready-to-(ab) use: From fake account trafficking to coordinated inauthentic behavior on Twitter. *Online Social Networks and Media*, 31:100–224.

Jesse McCrosky. 2020. How social media may redistribute trust away from institutions. *DataEthics Newsletter*.

Marshall McLuhan and Quentin Fiore. 1967. The medium is the message. *New York: Random House*, 123(1):126–128.

Marshall McLuhan and Bruce R Powers. 1989. *The global village: Transformations in world life and media in the 21st century*. Communication and Society.

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

Filippo Menczer. 2021. Facebook whistleblower Frances Haugen testified that the company's algorithms are dangerous — here's how they can manipulate you. *The Conversation*.

Marc Meola. 2004. Chucking the checklist: A contextual approach to teaching undergraduates web-site evaluation. *portal: Libraries and the Academy*, 4(3):331–344.

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

Miriam J Metzger. 2007. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091.

Miriam J Metzger and Andrew J Flanagin. 2015. Psychological approaches to credibility assessment online. *The Handbook of the Psychology of Communication Technology*, pages 445–466.

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

Bianka Młodkowska et al. 2019. Influencers on Instagram and YouTube and their impact on consumer behaviour. *Journal of Marketing and Consumer Behaviour in Emerging Markets*, 9(1):4–13.

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):180–212.

Rachel Moran, Izzi Grasso, and Kolina Koltai. 2022. Folk theories of avoiding content moderation: How vaccine-opposed influencers amplify vaccine opposition on Instagram. *Social Media + Society*.

Fred Morstatter, Jürgen Pfeffer, and Huan Liu. 2014. When is it biased? assessing the representativeness of Twitter's streaming API. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 555–556.

Eni Mustafaraj, Emma Lurie, and Claire Devine. 2020. The case for voter-centered audits of search engines during political elections. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 559–569.

Brigitte Naderer. 2023. Influencers as political agents? the potential of an unlikely source to motivate political action. *Communications*, 48(1):93–111.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence Survey Track*, pages 4551–4558.

National Geographic. 2009. United States Regions. `https://www.nationalgeographic.org/maps/united-states-regions/`.

TikTok Newsroom. 2020. New on TikTok: Introducing stitch. *Journal of Student Research*.

Ioanna Noula. 2018. I do want media literacy… and more. A response to Danah Boyd. *LSE Blogs*.

Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

Katherine Ognyanova, David Lazer, Ronald E Robertson, and Christo Wilson. 2020. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*.

Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on Amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419.

Michal Pecanek. 2021. Six important insights about title tags (953,276 pages studied). *Ahrefs Blog*.

Andrew Peck. 2020. A problem of amplification: Folklore and fake news in the age of social media. *Journal of American Folklore*, 133(529):329–351.

R Peeters and T Pulls. 2015. Regaining the end-users' trust with transparency-enhancing tools. *Dostupno: http://www. project-opacity. com*, 13(5).

Gordon Pennycook, Jabin Binnendyk, Christie Newton, and David G Rand. 2021. A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, 7(1):25293.

Gordon Pennycook and David G Rand. 2021a. Examining false beliefs about voter fraud in the wake of the 2020 presidential election. *The Harvard Kennedy School Misinformation Review*.

Gordon Pennycook and David G Rand. 2021b. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5):388–402.

Billy Perrigo. 2023. Why your Twitter feed is suddenly full of people you don't follow. *Time Magazine*.

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):3–25.

Jay Peters. 2023. Google is going to let you annotate search results. *The Verge*.

W James Potter. 2013. Review of literature on media literacy. *Sociology Compass*, 7(6):417–435.

Marc Prensky. 2001. Digital natives, digital immigrants part 2: Do they really think differently? *On the Horizon*, 9(6):1–6.

Nicolas Pröllochs. 2022. Community-based fact-checking on Twitter's Birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 794–805.

Mickey Rapkin. 2017. The social media platform that has Gen Z obsessed. *Wall Street Journal*.

David N Rapp and Nikita A Salovich. 2018. Can't we just disregard fake news? the consequences of exposure to inaccurate information. *Policy Insights from the Behavioral and Brain Sciences*, 5(2):232–239.

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 *Proceedings of the Computational Journalism Conference*, volume 5.

Magdalena Riedl, Carsten Schwemmer, Sandra Ziewiecki, and Lisa M Ross. 2021. The rise of political influencers—perspectives on a trend towards meaningful content. *Frontiers in Communication*, 6:752656.

Soo Young Rieh and David R Danielson. 2007. Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology*.

Julio Rieis, Fabrício de Souza, Pedro Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 357–366.

Horst WJ Rittel and Melvin M Webber. 1973. Dilemmas in a general theory of planning. *Policy Sciences*, 4(2):155–169.

Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22.

Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2021. The impact of fake news on social media and its influence on health during the covid-19 pandemic: A systematic review. *Journal of Public Health*, pages 1–10.

Daniel Röchert, Gautam Kishore Shahi, German Neubaum, Björn Ross, and Stefan Stieglitz. 2021. The networked context of Covid-19 misinformation: informational homogeneity on YouTube at the beginning of the pandemic. *Online Social Networks and Media*, 26:100164.

Ashley Rodriguez. 2018. YouTube's algorithms can drag you down a rabbit hole of conspiracies, researcher finds. *Quartz*.

Richard Rogers. 2013. *Digital methods*. MIT press.

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

Robert M Ross, David G Rand, and Gordon Pennycook. 2021. Beyond "fake news"': Analytic thinking and the detection of false and hyperpartisan news headlines. *Judgment and Decision Making*, 16(2):484–504.

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*.

Jörgen Sandberg and Haridimos Tsoukas. 2015. Making sense of the sensemaking perspective: Its constituents, limitations, and opportunities for further development. *Journal of organizational behavior*, 36(S1):S6–S32.

Joshua M Scacco and Ashley Muddiman. 2016. Investigating the influence of "clickbait" news headlines. *Engaging News Project Report*.

Joseph S Schafer and Kate Starbird. 2023. Post-spotlight posts: The impact of sudden social media attention on account behavior. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 273–277.

Norbert Schwarz, Lawrence J Sanna, Ian Skurnik, and Carolyn Yoon. 2007. Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology*, 39:127–161.

Sylvia Scribner. 1984. Literacy in three metaphors. *American Journal of Education*, 93(1):6–21.

SerpApi. 2020. Serpapi: Google Search API.

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*, pages 745–750.

Elisa Shearer. 2021. More than eight-in-ten Americans get news from digital devices. *Pew Research Center*.

Ryan P. Shepherd. 2020. Gaming Reddit's algorithm: r/the_donald, amplification, and the rhetoric of sorting. *Computers and Composition*, 56:102572.

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*, pages 324–339.

Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146:102551.

Jakub Simko, Matus Tomlein, Branislav Pecher, Robert Moro, Ivan Srba, Elena Stefancova, Andrea Hrckova, Michal Kompan, Juraj Podrouzek, and Maria Bielikova. 2021. Towards continuous automatic audits of social media adaptive behavior and its role in misinformation spreading. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 411–414.

Jay Smooth. 2018. History of media literacy, part 1: Crash course media literacy 2.

Brian G Southwell, J Scott Babwah Brennen, Ryan Paquin, Vanessa Boudewyns, and Jing Zeng. 2022. Defining and measuring scientific misinformation. *The Annals of the American Academy of Political and Social Science*, 700(1):98–111.

Brian G Southwell, Emily A Thorson, and Laura Sheble. 2017. The persistence and peril of misinformation defining what truth means and deciphering how human brains verify information are some of the challenges to battling widespread falsehoods. *American Scientist*, 105(6):372–375.

Lauren Southwick, Sharath C Guntuku, Elissa V Klinger, Emily Seltzer, Haley J McCalpin, and Raina M Merchant. 2021. Characterizing Covid-19 content posted to TikTok: public sentiment and response during the first phase of the Covid-19 pandemic. *Journal of Adolescent Health*, 69(2):234–241.

Seth Spaulding. 1966. The unesco world literacy program: A new strategy that may work. *Adult Education*, 16(2):70–84.

Jörg L Spenkuch and David Toniatti. 2016. Political advertising and election outcomes. *Kilts Center for Marketing at Chicago Booth–Nielsen Dataset Paper Series*, pages 1–046.

Isaac Stanley-Becker. 2020. Google greenlights ads with 'blatant disinformation' about voting by mail. *The Washington Post*.

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*, pages 79–111.

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):1–26.

Kate Starbird, Emma S Spiro, and Kolina Koltai. 2020. Misinformation, crisis, and public health—reviewing the literature v1. *Social Science Research Council, MediaWell*, 25.

StatCounter. 2021. Search engine market share worldwide. `https://gs.statcounter.com/search-engine-market-share`.

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

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.

Leo G Stewart, Ahmer Arif, and Kate Starbird. 2018. Examining trolls and polarization with a retweet network. In *Proceedings of the ACM WSDM Workshop on Misinformation and Misbehavior Mining on the Web*, volume 70.

S. Shyam Sundar, Maria D Molina, and Eugene Cho. 2021. Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, 26(6):301–319.

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*, 9(3):286–299.

Justina Tam, Emily K Porter, and Una J Lee. 2022. Examination of information and misinformation about urinary tract infections on TikTok and YouTube. *Urology*, 168:35–40.

Percy H. Tannenbaum. 1953. The effect of headlines on the interpretation of news stories. *Journalism Quarterly*, 30(2):189–197.

Kat Tenbarge. 2020. YouTube channels made money off of fake election results livestreams with thousands of viewers. *Insider.com*.

Bennie Thompson, Liz Cheney, and Zoe Lofgren. 2022. Thompson, Cheney, & Lofgren opening statements at Select Committee Hearing. *Social Media + Society*.

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

Daniel Trielli and Nicholas Diakopoulos. 2019. Search as news curator: The role of Google in shaping

attention to news information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*, pages 1–15.

Daniel Trielli and Nicholas Diakopoulos. 2022. Partisan search behavior and google results in the 2018 us midterm elections. volume 25, pages 145–161.

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):520–535.

Aleksandra Urman, Mykola Makhortykh, Roberto Ulloa, and Juhi Kulshrestha. 2022. Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and informatics*, 72:101860.

US Census Bureau. 2010. 2010 Census urban and rural classification and urban area criteria.

Shannon Vallor. 2016. *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

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*, volume 11.

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

Emily K Vraga and Leticia Bode. 2021. Addressing Covid-19 misinformation on social media preemptively and responsively. *Emerging Infectious Diseases*, 27(2):396.

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):632–652.

Claire Wardle. 2018. How we all can fight misinformation. *Harward Business Review*.

Jen Weedon, William Nuland, and Alex Stamos. 2017. Information operations and facebook. *Retrieved from Facebook News Room Files*.

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):199–206.

S White and M McCloskey. 2003. Framework for the 2003 national assessment of adult literacy (nces 2005-531). *US Department of Education. Washington, DC: National Center for Education Statistics*.

Joe Whittaker, Seán Looney, Alastair Reed, Fabio Votta, et al. 2021. Recommender systems and the amplification of extremist content. *Internet Policy Review*, 10(2).

Tom Wilson. 2021. *Understanding the Structure and Dynamics of Multi-platform Information Operations*. Ph.D. thesis, University of Washington Libraries.

Tom Wilson and Kate Starbird. 2020. Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School Misinformation Review*, 1(1).

Sam Wineburg and Sarah McGrew. 2017. Lateral reading: Reading less and learning more when evaluating digital information. *Stanford History Education Group Working Paper*.

Justine Wise. 2019. Twitter suspends account that helped incident with native american man go viral. *The Hill*.

Chloe Wittenberg, Ben M Tappin, Adam J Berinsky, and David G Rand. 2021. The (minimal) persuasive advantage of political video over text. *Proceedings of the National Academy of Sciences*, 118(47):e2114388118.

Samuel C Woolley and Philip Howard. 2017. Computational propaganda worldwide: Executive summary. *The Computational Propaganda Project*.

Samuel C Woolley and Philip N Howard. 2016. Political communication, computational propaganda, and autonomous agents: Introduction. *International journal of Communication*, 10.

Jing Wu and Yu-mei Wang. 2011. Unpacking new media literacy. *Journal of Systemics, Cybernetics and Informatics*, 9(2):84–88.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90.

Luh Putu Ayu Wulandari and Gede Sri Darma. 2020. Advertising effectiveness in purchasing decision on Instagram. *Journal of Business on Hospitality and Tourism*, 6(2):381–389.

Himanshu Zade, Morgan Wack, and Yuanrui Zhang. 2022a. Google's search headlines from 2020 US election: Data files.

Himanshu Zade, Morgan Wack, Yuanrui Zhang, Kate Starbird, Ryan Calo, Jason Young, and Jevin D West. 2022b. Auditing Google's search headlines as a potential gateway to misleading content: Evidence from the 2020 US election. *Journal of Online Trust and Safety*, 1(4).

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. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–27.

Brandy Zadrozny and Ben Colins. 2020. Antifa rumors spread on local social media with no evidence. *NBC News*.

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation warfare: Understanding State-sponsored trolls on Twitter and their influence on the web. In *Companion Proceedings of the 2019 World Wide Web Conference*, pages 218–226.

Eric Zeng, Miranda Wei, Theo Gregersen, Tadayoshi Kohno, and Franziska Roesner. 2021. Polls, clickbait, and commemorative $2 bills: problematic political advertising on news and media websites around the 2020 US elections. In *Proceedings of the 21st ACM Internet Measurement Conference*, pages 507–525.

Jing Zeng, Mike S Schäfer, and Joachim Allgaier. 2020. Reposting "till Albert Einstein is TikTok famous"': The memetic construction of science on TikTok. *International Journal of Communication*, 15:3216–3247.

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

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

Erfei Zhao, Qiao Wu, Eileen M Crimmins, and Jennifer A Ailshire. 2020. Media trust and infection mitigating behaviours during the Covid-19 pandemic in the USA. *BMJ Global Health*, 5(10):e003323.

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):102390.

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

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*, pages 493–502.

Matthew A Zook and Mark Graham. 2007. Mapping digiplace: geocoded internet data and the representation of place. *Environment and Planning B: Planning and Design*, 34(3):466–482.