University of Oxford

Managing Online Rumour Proportions During Offline Protests

Cameron Raymond

Thesis submitted in partial fulfilment of the requirement for the degree of MSc in Social Data Science at the Oxford Internet Institute at the University of Oxford

Trinity Term 2021

12,128 words

Acknowledgements

I would like to dedicate this thesis to my late grandpa, John Edward Keenan. I'm grateful he was able to see me begin my journey at Oxford, even if he never saw its completion. He was a champion of higher education, despite never being able to attend university himself, and I hope I made him proud.

Completing this degree during a global pandemic is entirely due to the grace and goodwill of others. I am extremely grateful to my supervisors, Dr Peaks Krafft and Dr Joss Wright for their advice and counsel. In addition to all the teaching staff, I would also like to thank my peers at the Oxford Internet Institute and Exeter College, who I am lucky to call friends. Finally, I'd like to thank my family — my mother, dad and brothers — and my partner Miranda for their unwavering support.

Abstract

Protests are a key mechanism for collective action and are increasingly reliant on online communication. In addition, large, offline protests are also events with tremendous amounts of ambiguity. Those on the ground are the first to know of new developments, but it is difficult to verify information in a timely manner. Furthermore, little experimental research has been done to understand how design factors online affect the information that activists have access to. This preregistered study investigates this gap by implementing an experimental analysis of rumour and misinformation sharing during a mock-protest demonstration. Through a Bayesian, "rumour proportions" framework, this study observes how adding credibility indicators (*i.e.*, warnings) to misleading content affects the belief in, and rate at which participants share, rumours. Additional exploratory analyses examine the role of post attributes in sharing and the possibility of an "implied truth effect." The results show that the treatment made participants share in greater alignment with what they perceived as accurate, but did not alter those perceptions themselves. Thus, those with biased accuracy estimates continued to share misleading information. By developing a custom experimental environment, moving beyond general misinformation stimuli and focusing on a specific application of credibility indicators, this study calls into question the notion that individuals spread misleading information simply because their attention is focused on factors other than accuracy.

Contents

1 Introduction

Protest demonstrations are a fundamental and extremely common form of political participation. Anti-racism protests in response to the murder of George Floyd peaked in magnitude on June 6, 2020 — with over 500,000 protestors taking to the street across 550 locations throughout the United States (Buchanan, Bui, & Patel, 2020). Such demonstrations propelled conversations of racial inequity, criminal justice and police brutality into the mainstream. In a recent poll, 68% of Americans said that the right to protest is very important for the country (Pew Research Center, 2020). By calling attention to inequities, protests are manifestations of dissent and express the need for political, cultural, or legal change.

Additionally, demonstrators are increasingly reliant on online communication, through social media platforms and direct messaging applications, before and during protests (Earl, McKee Hurwitz, Mejia Mesinas, Tolan, & Arlotti, 2013). Social media allows demonstrators and bystanders to broadcast information in real time, increasing the speed and reach of communication. Having access to high quality information on social media during demonstrations is especially important as such information is generally too timely, local, or hard to verify to be reported elsewhere. Thus, social media represents a key piece of infrastructure for this form of political participation.

Concurrent to the rise of social media have been concerns about its role in facilitating the spread of misinformation, and misinformation's resultant harms. The susceptibility of online platforms to misinformation has received ample attention in recent years. While the bulk of this scholarly attention has been paid to political and scientific misinformation, protest misinformation has the potential to sew confusion and conflict in what are, often, already charged environments (Trana, Valechaa, & Raoa, 2020). Thus, concerns over the ability for misinformation to undermine protest demonstrations are well warranted. Nonetheless, while protest misinformation has invariably been spread, that only characterizes what is likely a small portion of the information ecosystem during demonstrations. Given the ambiguity that can surround large protests, much of the information shared during demonstrations is at its core unverifiable and thus is best categorized as rumour. However, such ambiguity does not imply that unverified information shared online can be less plausible or pose fewer harms to those who believe and spread it. Accordingly, there are benefits for platforms to manage ambiguous and timely information in a transparent and community-driven manner.

Thus, with respect to offline activism social media presents a double-edged sword. Platforms serve as crucial communication infrastructure during protest demonstrations, but also as the means through which misleading content is spread. This tension is compounded by the intrinsic ambiguity of protests themselves. Developing interventions that improve the quality of information that activists have access to during demonstrations is the high-level objective of this research. This thesis contributes an empirical, causal understanding of the use of design factors to reduce the spread of misleading information in ambiguous contexts. This is done by designing interventions that encourage the thoughtful deliberation of social media posts.

To do so, this preregistered study examines how — given rumours with varying degrees of evidence — design interventions alter sharing behaviour in ambiguous contexts. The case of interest for this study are North American protest demonstrations. In addition, I present exploratory analyses which investigate the effects of source attributes and an implied truth effect. These analyses leverage a minimal clone of a Twitter-like platform I implemented from scratch for the purposes of conducting behavioural experiments, increasing ecological validity. Through fine-tuned control of the experimental environment, real-world stimuli scraped from Twitter, and carefully chosen attention checks I collect high quality behavioural data for analysis. When situated within a regression framework, the results present a nuanced picture with respect to the efficacy of credibility indicator interventions in rumour systems. The results show that the treatment made participants share in greater alignment with what they perceived as accurate, but did not alter those perceptions themselves. Thus, those with biased accuracy estimates continued to share misleading information. By moving beyond general misinformation stimuli and focusing on a specific application of design interventions, this study calls into question the notion that individuals spread misleading information simply because their attention is focused on factors other than accuracy.

This thesis begins by reviewing the relevant literature on protest demonstrations and social media; misinformation and its debiasing; and rumour systems. Then, Krafft and Spiro's "rumour proportions" framework is outlined, reconciling the relevant experimental literature on misinformation with ambiguous settings, in which information cannot be verified. This framework is operationalized into an experimental design in the methodology section. The results section, then, presents preregistered confirmatory analyses as well as preliminary exploratory analyses. Finally, the methodological, practical and theoretical implications of applying design interventions to rumour systems, and this study's limitations, are considered in the discussion section.

2 Literature Review

The diversity of information shared online, and the ends such information can be used for, imply that general solutions to misleading content are likely problematic. In order to contextualize this study's setting, *i.e.* protest demonstrations, this section first reviews the literature on social media usage during protests. While understanding the role of social media during protests is useful, it says little about how the structure and design of social media platforms affect the information activists have access to. I then review the relevant observational and experimental literature on misinformation and its debiasing. This acts as a helpful grounding for this study's experimental design; however, the literature's reliance on deterministic classifications of information veracity are ill-equipped to handle the ambiguity of protest scenarios. Thus, I finally look at the role of rumour in the broader information ecosystem, noting its importance in ambiguous contexts like protests and crisis situations. In tandem, these strands of literature provide a contextualized backdrop with which this study's theoretical framework and experimental design can be implemented.

2.1 Protest Demonstrations and Social Media

Protests are a key mechanism for collective action and are increasingly being organized online (Steinert-Threlkeld, Mocanu, Vespignani, & Fowler, 2015). Grassroots movements are able to take advantage of the broad reach that social media provides when mobilizing activists, sharing information about events, and discussing protest objectives. Alicia Garza, Patrisse Cullors-Brignac, and Opal Tometi helped co-found Black Lives Matter (BLM) and crystalized a movement for racial equity in the United States by using online platforms and the hashtag #BlackLivesMatter as foundational communication infrastructure (De Choudhury, Jhaver, Sugar, & Weber, 2016). While shown to be a useful agenda-setting tool (Freelon, McIlwain, & Clark, 2018) and important for the organization of movements and protests (Van Laer & Van Aelst, 2010), Earl et al. (2013) illustrate how online platforms are critical *during* demonstrations. For example, they found that during the 2009 G20 meetings in Pittsburgh, PA, social media allowed protestors to share the location of police — which previously only law enforcement would know — and notify others of police violence. Similar work studying the Gezi Park movement found trace data from Twitter to be predictive of offline political events (Varol, Ferrara, Ogan, Menczer, & Flammini, 2014). Given historic power imbalances between the state and activists, social media can be an important lever for reducing information asymmetries.

Using individual-level data from Belgian protests, Van Laer and Van Aelst (2010) parse out the differences between activists who are active online and those who are not. They find significant sociodemographic and political difference — with "online" activists being highly educated, experienced, and embedded in their movements. This follows a strand of literature which suggests that the internet can, counterintuitively, increase participation inequalities among activists. However, reduced communication costs and increasing access to the internet suggests that these inequalities may not be entrenched (Kelly Garrett, 2006). Given this, the focus on online communication during protests may not provide insight into offline communication, despite their necessary interrelation, due to population and medium differences.

Given the speed with which protest dynamics can change, the feedback loop between events and their coverage by traditional news outlets is often too long to be useful to activists during demonstrations. Social media reduces this feedback loop by allowing the real-time broadcasting of information by activists and bystanders themselves (Earl et al., 2013). Thus, the use of online platforms to alert demonstrators to protest developments is especially relevant. Social media allows for the rapid spread of information, unmediated by traditional news outlets; however, algorithmic mediation online means that this does not necessarily grant activists greater control over the information ecosystem (Poell & Van Dijck, 2015). While useful in temporally constrained situations, it also makes the platforms susceptible to misleading information. Using a comparative analysis of 30 protests occurring in the United States and Germany between 1960 and 2010, Nassauer (2018) find that a breakdown in communication is a common predecessor to the emergence of violence during protests.

While it is clear that online platforms are heavily used by activists before and during protests, and that having access to pertinent reliable information is important — little work in this realm has gone beyond descriptive observational studies. There is a lack of research focusing on how the structure and design of online platforms affect the quality of information that demonstrators have access to. This study aims to help build the evidence base for policy and design interventions that improve information systems in ambiguous contexts like protests.

2.2 Misinformation and its Debiasing

An area of work that has received ample attention in recent years surrounds misinformation, and attempts to prevent its belief and spread. Misinformation is defined as an unintentionally false utterance, while disinformation is deliberately false or misleading (Jack, 2017). Given the difficulties in assessing intentionality, this section will focus on misinformation. Efforts to reduce false information online can be grouped in to three strands: misinformation detection; user experimental studies, which propose and evaluate interventions at the user-level in reducing the spread or belief in false information; and observational studies, which characterize misinformation's spread in actual settings (Micallef, He, Kumar, Ahamad, & Memon, 2020). This study proposes an experimental approach, but from both a practical and theoretical standpoint the entire misinformation literature is important and relevant to consider.

Misinformation detection. For misinformation to be stemmed, it must first be identified. In various contexts, automated approaches to misinformation detection that model user, source, and post features have been proposed (Wang et al., 2018; Islam, Liu, Wang, & Xu, 2020; Sitaula, Mohan, Grygiel, Zhou, & Zafarani, 2020). Additionally, given that misinformation is frequently created on third-party websites and then shared on social media, problematic news articles can be identified by verified fact-checking organizations (Friggeri, Adamic, Eckles, & Cheng, 2014). This approach, however, inevitably allows posts that reference a piece of misinformation, but do not link to the misinformation's source, to go undetected. While URL matching or algorithmic detection approaches may work in some cases, they are unlikely to generalize to the smaller, local, ambiguous, and temporally constrained pieces of information that are shared and relied upon during protests. Sharma et al. (2019) distinguish between content-, feedback-, and intervention-based detection methods for misinformation. Unlike content-based approaches, which focus on the content of social media posts, feedback-based methods rely on user responses to identify misleading information. While recent research has shown that crowdsourcing accuracy judgments of news sources can provide unbiased estimates, it is unclear if this approach would generalize to other contexts (Pennycook & Rand, 2019; Allen, Arechar, Pennycook, & Rand, 2020). As will be seen, this poses a practical limitation to interventions in ambiguous contexts.

User experimental studies. In most cases, misinformation detection is not an end itself. Rather, it is generally a precursor to some intervention. While detected misinformation can be algorithmically depressed — user-interventions that warn individuals about misleading content can be more transparent and thus are especially pertinent in ambiguous contexts. Such interventions are most frequently implemented through credibility indicators. Credibility indicators are user interface

(UI) elements such as icons and text that provide cues regarding the credibility of a piece of information (Yaqub, Kakhidze, Brockman, Memon, & Patil, 2020). An experimental framework allows for randomization and strict control over the stimuli and intervention — allowing for stronger, causal claims than observational studies can provide.

One objective of misinformation interventions is to prevent the *belief* in misleading information. An influential study by Ecker, Lewandowsky, and Tang (2010) found that warning participants that a statement was false only partially eliminated participants belief in the statement. This is an example of the continued influence effect (CIE), where individuals rely on false information despite knowing that it is false (Johnson & Seifert, 1994). Similar complications were found by Pennycook, Bear, Collins, and Rand (2020), who observed an "implied truth effect." This is an unintended consequence of attaching credibility indicators to a *subset* of false posts where the absence of a credibility indicator was considered to imply truthfulness, even when it is not warranted. Additional work found adding stance labels, which indicate ideological affiliation, to news articles intensified selectivity bias (Gao, Xiao, Karahalios, & Fu, 2018). Finally, in an overview of the cognitive biases that prevent misinformation's debiasing Lewandowsky, Ecker, Seifert, Schwarz, and Cook (2012) identify four key factors that need to be addressed: CIE; a familiarity backfire effect, where repeating misinformation increases familiarity, reinforcing it; an overkill backfire effect, where complex refutations are cognitively burdensome relative to simple myths; and a worldview backfire effect, where discordant information strengthens prior beliefs. In addressing these, the authors cite five relevant design considerations: to provide a contrastive, alternative account of events; to emphasize facts over the original myth; to, whenever possible, add preexposure warnings to misleading information; to provide simple and brief rebuttals to misinformation; and to foster healthy skepticism about information. Such cognitive and human-computer interaction concerns are crucial in forming effective interventions.

While the body of evidence on credibility indicators suggests they are generally useful in correcting the belief of misleading information, only recently have studies begun to examine their effect on *sharing intentions*. Sharing intentions can act independent of accuracy judgements because the decision to reshare information on social media is multi-faceted, with accuracy being one factor (Pennycook et al., 2021). Other considerations are whether the information is surprising, politically concordant, funny, and interesting (Pennycook et al., 2021). Two studies by Pennycook, McPhetres, Zhang, Lu, and Rand (2020) focusing on the COVID-19 pandemic found that individuals were better at discerning true and false content when explicitly asked about its accuracy as opposed to their intent to share the content on social media. Thus, an argument for why effective credibility indicators work is that they increase the saliency of accuracy in the sharing calculus (Pennycook et al., 2021). This advances a "cognitive-inattention" based account of misinformation sharing over one of ideological motivated reasoning, whereby people are biased against believing information that challenges their political ideology (Garrett & Weeks, 2013). Pennycook et al.'s work has since been replicated (Roozenbeek, Freeman, & van der Linden, 2021); however, the necessary conditions that are conducive to cognitive-inattention based interventions has not been explored.

In addition, different scenarios may not allow for professional fact-checking, as would be the case with most ambiguous contexts. Yaqub et al. (2020) evaluate the impact of credibility indicators that cite different ground truths on participants intent to share content. Depending on whether the credibility indicator was said to be based on the judgement of professional fact-checkers, the news media, artificial intelligence (AI) or majority opinion they found varying effect sizes by political affiliation, age, gender and social media usage. Similar research focusing on the Singaporean case has found that media-skepticism is an additional moderating factor (Nekmat, 2020).

Observational Studies. Finally, observational studies seek to characterize misinformation on social media platforms. This includes misinformation's prevalence, amplification, reach, diffusion, and organic correction. Friggeri et al.'s paper *Rumour Cascades* (2014) takes a social network analysis approach to misinformation, forming a quasi-experimental analysis of the effect of "snoping"[1] on rumour propagation. Contrary to Mosleh, Martel, Eckles, and Rand's (2021) field experiment, which found adverse downstream effects to organic corrections, they found that linking to a fact-checking website reduced the spread of the post being snoped (Friggeri et al., 2014). In a similar vain, Micallef et al. (2020) analyzed COVID-19 conspiracy theories on Twitter; the properties of those posting, and countering, misinformation; and the effect of counter-misinformation. They found that given the amount of COVID-19 misinformation being produced, professional fact-checkers currently have a limited effect. However, they also found that misleading information was quick to be corrected by other users on Twitter. It is of note that all the studies discussed relied on fact-checking websites or traditional media sources as their source of misinformation and ground truth evaluations. Rumours that are evaluated by professional fact-checking organizations are generally larger, and thus may not gen-

[1]Snoping is defined as linking to the fact checking website www.snopes.com in response to a rumour.

eralize to smaller, more local and more temporal pieces of information that would characterize protests, posing practical limitations.

When searching for insights as to how best counter misinformation, it is useful to turn to the social media platforms themselves and observe how communities are already doing so organically. Social media users already link to trusted sources and provide alternative explanations for events without platform intervention. Basing interventions on existing practices helps ensure their utility and institutionalizes positive norms.

2.3 Rumour Systems

It is clear that online communication plays an important role during protests, and that online platforms have taken on a crucial function as a broadcasting tool. Given the harms that result from communication breakdowns, a key objective would be to reduce the proliferation and belief in misleading information. While the relevant misinformation literature can provide insights as to possible interventions, a focus on misinformation is too narrow. This is because much of the information shared during protests is unverifiable and is therefore rumour. Indeed, an analysis by Brennen, Simon, Howard, and Nielsen (2020) found that 59% of COVID-19 "misinformation" contains true information that is spun, twisted, recontextualized, or reworked — complicating a simple true or false binary. Rumour, separate from misinformation, is best categorized as "unverified information" where the proposition's veracity is unknowable (Shibutani, 1966). Thus, it is only in retrospect that a rumour can be labelled as true, false, misinformation, or disinformation (Bordia & DiFonzo, 2004). Social media amplifies the capacity for rumours to spread, increasing their scholarly relevance and, of practical importance, provides novel methods for investigation.

Rumour is an important facet of ambiguous social situations, of which protests are but one important example. Focusing on crisis situations, Huang, Starbird, Orand, Stanek, and Pedersen (2015) interviewed those who used social media in the aftermath of the 2013 Boston Marathon Bombings. They found that the speed at which information is shared on social media and its decentralized nature are distinguishing factors — making social media especially useful in crisis situations — but also more susceptible to misinformation, intentional or otherwise. Rumour systems contain spatial and temporal elements. Spatial, temporal and emotional proximity are all associated with rumouring during crises (Huang et al., 2015; Oh, Agrawal, & Rao, 2013). Without credible sources to turn to, Huang et al.'s interviewees turned to source attributes, like if they know or trust the individual sharing the information, or content attributes to determine information's reliability.

Given the intrinsic ambiguity of rumours, it may be tempting to try and stem them all together. Yet, this would be a mistake. Either because the situation is too ambiguous, local, or temporally constrained — rumours are sometimes all that are available. Rumours introduce new discourses into the information ecosystem and many events that are later reported on as news are first reported by individuals as rumours, a process which Shibutani (1966) refers to as "improvised news." Thus, rather than simple gossip — rumours are a powerful mechanism of communication. Time series analyses of rumour systems (Ma, Gao, Wei, Lu, & Wong, 2015) and rumour stance classifications (McCreadie, Macdonald, & Ounis, 2015; Zeng, Starbird, & Spiro, 2016) provide empirical evidence that rumouring acts as a form of collective sensemaking in uncertain environments. As this takes place, Starbird et al. (2016) note a shift from speculative rumouring to presenting rumours as factual. A recent study focusing on the case of India's demonetization policy, which severely disrupted existing market practices, argues that rumour is a mechanism through which disenfranchised communities can "collectively bond and make sense of their own place in society" (Chandra & Pal, 2019). Similarly, others have argued that informal means of communication can be an integral tool for the preservation of knowledge and safety against oppressive institutions (Fine & Turner, 2001). Thus, there is a distinctive political dimension to rumour as well.

To view rumours as a problem to be solved is incomplete. While problematic if given more credence than is reasonable, rumours are an important mechanism for communication when alternatives do not exist or are not desirable. Thus, rather than eliminating rumours, interventions ought to encourage the thoughtful deliberation that allows for collective sensemaking.

3 Theoretical Framework

In tandem, the previous literature on the use of online communication during protests, experimental interventions on misinformation, and the role of rumouring in uncertain environments paint a complicated picture. If implemented properly — that is, in a social, democratic and community-driven manner — there are likely benefits to managing implausible information during demonstrations. However, the focus previous experimental research has had on verifiable information is ill-equipped to map to rumour systems.

As will be argued, this complication can be remedied through a Bayesian approach to information management. To do so, this paper adopts the framework introduced by Krafft and Spiro (2019). They outline a system-level framework for managing uncertain information. Rather than stifling unverified information entirely, the framework aims to encourage healthy levels of "rumour proportions," relative to normative benchmark representations of intrinsic uncertainty (Krafft & Spiro, 2019). This expands the study of misinformation beyond the special cases of true or false claims to include the superset of all information, verified or otherwise.

To do so they advocate for a framework that focuses on rumour proportions, in which belief in a rumour is compared with the plausibility of that rumour's proposition. They define a rumour R as a logical proposition about the world. Within the target population, say a hypothetical social media platform, N_R then is the number of people or posts discussing R, of which n_R affirm a belief in R. They define the rumour proportion of R then to be $p_R = \frac{n_R}{N_R}$. They then define three design criteria that attempt to contextualize R with respect to its evidence.

1. If rumour proportion p_R is at a "reasonable" level, then new evidence for R should increase p_R.

2. If rumour proportion p_R is at a "reasonable" level, then new evidence against rumour R should decrease p_R.

3. If rumour $R1$ is more plausible than rumour $R2$ according to publicly available evidence, and both rumour proportions are at "reasonable" levels, then we ought to observe $p_{R1} > p_{R2}$ (Krafft & Spiro, 2019, p. 5).

Evidence levels E for R are quantified adopting a Bayesian approach, and related to the rumour proportion benchmark such that $f\colon P(R|E) \mapsto p_R$ in a monotonically increasing manner. For example, let us define $f(P(R|E)) = p_R$ and some rumour R which, given known evidence E, has a 60% likelihood of being true ($P(R|E) = 0.6$). In this scenario, the normative benchmark for the system's affirmation rate – that is, the percentage of posts in agreement with the rumour – ought to be 60% ($p_R = 0.6$). Thus, Krafft and Spiro provide a descriptive framework for how rational systems might look, and a normative framework that prescribes what ideal rumour proportions may be in a given context.

Krafft and Spiro's framework provides the basis to expand the experimental work done on misinformation to include other forms of unverified information. More importantly, it helps researchers move beyond a true/false binary that, as previously discussed, is problematic. As previously noted, research on rumour by Friggeri et al. (2014) collected Facebook threads discussing rumours and observed the effect of a comment linking to a fact-checking website on subsequent engagement. The fact-checking website, Snopes, will label propositions as true, false, partly true, unknown or satire; however, the researchers predominantly analyzed true or false propositions, bucketing the remaining labels and mostly ignoring them (Friggeri et al., 2014). This conflates various forms of information that are relevant facets of the public's media ecosystem, and is common in misinformation research. Within a rumour proportions framework, true information and misinformation become special cases of proposition where $P(R|E) \approx 1$ and $P(R|E) \approx 0$ respectively. Thus, the totality of the information ecosystem, including ambiguous propositions which before were necessarily omitted, can now be incorporated and analyzed.

While providing more nuance to the study of information systems, a rumour proportions approach carries stronger assumptions that must be managed. This is most notable when quantifying the evidence level for a rumour, and carries practical and epistemic concerns.

First, in taking on a rumour proportions framework, researchers, content moderators, community members, and practitioners must grapple with estimating the evidence levels of propositions. Practically, this is much more difficult than restricting the analysis to information that is verifiably true or false and performing a binary classification. While it may not be feasible to directly evaluate $P(R|E)$ for a rumour, this can be avoided by comparing pairs of rumours. Thus, as long as it is possible to determine that $P(R1|E) > P(R2|E)$ one can still assess whether the rumour proportions are balanced or not (Krafft & Spiro, 2019).

Second, quantifying evidence levels carries epistemic and political weight. This is because in many cases, there may be reasonable debate over the plausibility of a rumour and what the evidence level ought to be. Thus, whose perspective is given credence in this calculus is pivotal. Yet, limiting the focus of research to a binary classification does not eliminate these value laden decisions, as one still has to determine what narratives to elevate in determining propositions' veracity (Marwick, Kuo, Cameron, & Weigel, 2021). Instead, it focuses attention on "low-hanging fruit" where expert consensus can be achieved and agreed upon. Thus, a deterministic approach to misinformation does not alleviate these concerns. Rather, it can serve to obscure these important decisions as "common sense." Conversely, the political dimension to misinformation is particularly acute when analyzing politically charged rumours or information that is shared within networks of marginalized groups (Marwick et al., 2021). Given that these are decisions that all relevant stakeholders must grapple with, it is beneficial to make such conversations public. Making value laden decisions, like what evidence to incorporate into the rumour proportion calculus, transparent is consequently a necessary first step.

3.1 Experimental Extension

Krafft and Spiro introduce a more nuanced view of information systems, presenting opportunities to extend existing work on misinformation interventions. The following experimental design extends the existing experimental literature of misinformation to encompass not just true or false information, but all propositions with varying levels of plausibility. It is only able to do this through Krafft and Spiro's Bayesian approach to rumour systems. As noted earlier, an important example of where such an approach may be useful is the case of protest demonstrations. This thesis utilizes rumour proportions as the normative benchmark and grounding for an online social media experiment where participants share rumours of varying evidence levels that would be relevant in a protest scenario.

4 Methodology

As discussed in the previous section, Krafft and Spiro's conception of rumour proportions provides a nuanced normative framework for the analysis of information online. A rumour proportions framework is especially useful in situations where the veracity of information cannot be assessed deterministically, as is often the case during protest demonstrations. The following experimental design assesses the efficacy of crowdsourced credibility indicators in nudging rumour systems into proportion. This is done through a mock-protest environment where participants are asked to share posts discussing two rumours: that protesters are being kidnapped by federal agents ($R1$) and that law enforcement are using digital contact tracing technology to track protestors ($R2$).

All materials and plans for analysis were preregistered on the Open Science Framework (see, https://osf.io/qrksg).

4.1 Confirmatory Hypotheses

This study applies a rumour proportions framework to a mock-protest rumour system online through the introduction of credibility indicators. Given the limitations of focusing on verifiable misinformation, this study seeks to evaluate credibility indicators' efficacy in an ambiguous environment and to encourage thoughtful sharing of information, without stemming the utility of rumours entirely. However, resharing information related to a rumour is not equivalent to affirming a belief in that rumour. As such, I adopt a common qualitative coding of rumour into four categories: affirmations, denials, neutral statements, and questions (Krafft & Spiro, 2019). Affirmations and denials are of particular interest and form the basis of the study's four confirmatory hypotheses.

- H_1: The presence of credibility indicators in the system will decrease the affirmation rate for rumours with lower evidence levels.
- H_2: The presence of credibility indicators in the system will increase the denial rate for rumours with lower evidence levels.
- H_3: The presence of credibility indicators in the system will increase the affirmation rate for rumours with higher evidence levels.
- H_4: The presence of credibility indicators in the system will decrease the denial rate for rumours with higher evidence levels.

4.2 Experimental Design

As noted previously, one can avoid specifying evidence levels for rumours by comparing them ordinally. This study, then, will limit itself to two rumour types: high evidence rumours and low evidence rumours. For simplicity, this study will take on a 2×2 mixed factorial design, using the presence of credibility indicators as a between-subjects factor (condition assignment: credibility indicators vs. no credibility indicators) and the evidence level of various rumours as a within-subjects factor (evidence level: high vs. low evidence).

With respect to H_1–H_4, there are two dependent variables of interest: the rate at which participants affirm a rumour and the rate at which participants deny a rumour. We can define the affirmation rate (p_R^a) for a rumour (R) as the number of posts that a participant shares affirming R (n_R^a) divided by the total number of posts affirming R (N_R^a). The denial rate (p_R^d) for R can be defined similarly as $p_R^d = \frac{n_R^d}{N_R^d}$. If $P(R1|E) > P(R2|E)$ according to publicly available evidence, and both rumor proportions are at "reasonable" levels, then we ought to observe $p_{R1}^a > p_{R2}^a$ and $p_{R1}^d < p_{R2}^d$ (Krafft & Spiro, 2019).

4.3 Stimuli

The stimuli for this experiment are social media posts and videos derived from Twitter during recent anti-racism protests in the United States. Videos are used to contextualize the social media posts that participants interact with.

The first set of stimuli pertains to $R1$, that protestors were being "kidnapped" by federal agents without identifying themselves. This rumour originated from a video posted on Twitter showing a protestor being taken and driven away in an unmarked van. According to the fact-checking website Snopes, no court or judge determined that any protesters were "kidnapped" by federal agents as this implies a crime (Lee, 2020). However, it is true that protesters were being detained in unmarked vans, and in multiple cases, those who were detained claimed that this occurred without the agents' identifying themselves. As such, Snopes labelled this rumour as a mixture of true and false (Lee, 2020). Broadcasting state violence is an important function of social media during demonstrations (Earl et al., 2013). Thus, $R1$ is defined as a high evidence rumour as it has been supported by third party sources and, legal semantics aside, is highly relevant to demonstrators.

The second set of stimuli pertains to $R2$, that protestors were being tracked by police using contact tracing technology originally intended for public health purposes. $R2$ originated from a video posted on Twitter where a city official used the term "contact tracing" as a metaphor for uncovering any links to organized crime that arrested demonstrators may have had. While spurring fear that law enforcement had access to critical public health data needed to stem the COVID-19 pandemic, this was not the case. No evidence was found in support of $R2$ and the rumour was later refuted by major news outlets as a poorly worded and misinterpreted metaphor (Morrison, 2020). Given this, and the harms that could come from a loss of faith in public health measures among demonstrators, $R2$ will serve as a low evidence rumour.

Importantly, both $R1$ and $R2$ were not verifiable within the relevant timeframe of an ongoing protest. However, $R1$ is much closer to the truth, and far more relevant to protesters, than $R2$ which is unequivocally false. Using third party fact-checking websites and news outlets as a ground truth is common practice in the study of misinformation, and for the purposes of this study are appropriate as well (Yaqub et al., 2020; Nekmat, 2020). However, in practice, neither demonstrators or content moderators would have the liberty of such fact-checking websites (Huang et al., 2015). In addition, while both rumours are from recent anti-racism protests, neither video contains any references to a specific social movement. Thus, these rumours serve as useful general protest stimuli.

As previously noted, both rumours originated from videos posted on Twitter, with each "thread" generating a large number of user replies and discussion. Using the Twitter application programming interface (API), I collected all of the replies for $R1$ ($N = 251$) and $R2$ ($N = 1293$) in the 24 hours after each video was initially posted, resulting in a total of 1544 replies. After collecting the Twitter thread for each rumour, I coded each reply according to whether it affirms, denies, is neutral

		Overall	*Rumour:* Kidnapping Rumour	Contact Tracing Rumour
N		167	86	81
Code	Affirms	70	23	47
	Denies	46	32	14
	Neutral	33	19	14
	Questions	18	12	6

Table 1: Summary of coded replies for $R1$, the rumour that police were “kidnapping” protesters (high evidence), and $R2$, that law enforcement are using contact tracing technology to track protesters (low evidence).

towards, or questions the rumour, with a fifth category for irrelevant replies or replies that would not have sufficient context to be understood by participants. The full codebook can be seen in Appendix A. After coding the rumour dataset twice, with a week in between each coding session and blind to previous codes, I observed a high degree of internal consistency (Cohen’s $\kappa = 0.768$). Conflicting codes were resolved in a third coding session and replies coded as irrelevant ($N = 1377$) were discarded – leaving 167 remaining tweets to be used in the experimental environment. These replies serve as the social media posts that participants interact with in the mock social media environment.

As can be seen in Table 1, which shows the number of replies assigned to each category, the proportion of replies denying $R1$ is much greater than $R2$, and the inverse is true for the proportion of replies affirming. Thus, according to Krafft and Spiro, this system is out of proportion, posing clear harms.

4.4 Procedure

The procedure for this study has four components: the preliminary questionnaire, the social media protest scenario for $R1$, the social media protest scenario for $R2$, and the post-study questionnaire. The following procedure was validated for ease and clarity by nine graduate students before use. The experimental environment can be viewed online (see, https://pedantic-bohr-7fc231.netlify.app/).

Upon accessing the experiment’s webpage, all participants are randomly assigned to one of two conditions (treatment vs. control) and asked to complete a preliminary questionnaire. This measures basic demographic information including age, gender and education — as well as left-right political affiliation, personal involvement in activism, what social movements they affiliate with and what social media

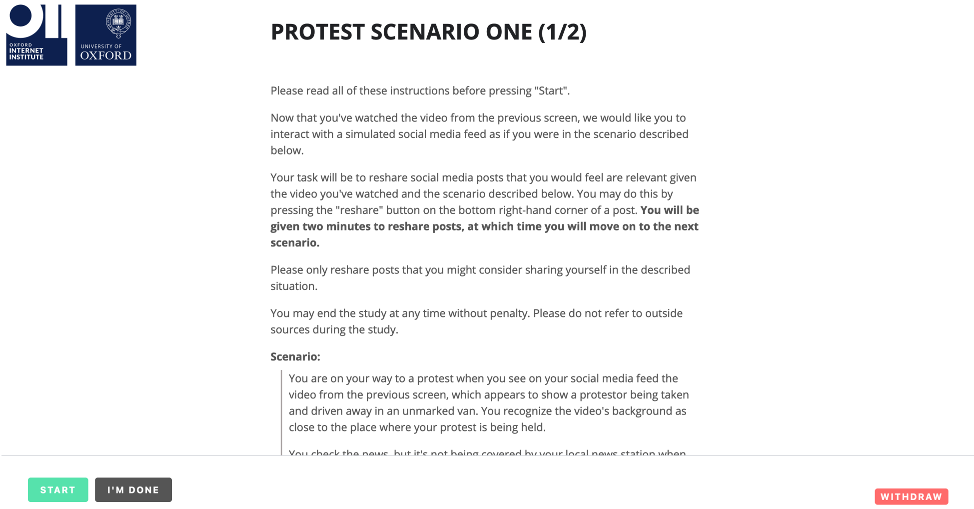

Figure 1: Staging webpage prior to starting experiment.

platforms they use. This serves two purposes. First, it will allow for the exclusion of participants who do not use social media or do not attend demonstrations. These are individuals who are unlikely to contribute to the misinformed spread of rumours and are outside the scope of the relevant population (Amaya, Biemer, & Kinyon, 2020). Second, these are necessary controls when analyzing the experiment's results. Given that online recruitment platforms rarely provide representative samples of the defined population, it is best practice to control for these demographic attributes (Coppock, 2019). Additionally, while this study does not aim to study the movements underlying protests, it is important to be cognizant that various movements may use social media platforms differently and thus control for different movement affiliations.

After completing the preliminary questionnaire, participants are asked to watch the video, downloaded from Twitter, that catalyzed $R1$ of a protestor being illegally detained and driven away in an unmarked van. Given that rumours are highly contextual, and thus hard to evaluate in isolation, this provides important background information to the rumour (Rosnow, 1988). Given the sensitive and potentially disturbing nature of the video, a content warning is given and the participant is reminded that if they expect this content to cause them significant distress that they should withdraw from the experiment. After watching the video, the participant is given a multiple-choice question asking what the video showed, and a question asking if they had seen the video before. The former question serves as an attention check to ensure that they watched the video. If the participant indicates being aware of the rumour, then it may be the case that their prior evidence level differs from what the stimuli suggests. As such, a participant indicating their awareness of either rumour, or failing an attention check will be used as further exclusion criteria.

After watching the video that catalyzed $R1$ the participant enters a staging page prior to entering the social media environment. Participants are told that their task is to reshare social media posts that they would feel are relevant for other demonstrators to be aware of, given the protest scenario. Thus, rather than simply measuring belief in rumours, this study also measures participants' propensity to sharing different forms of information given different design constraints. Participants are also given the short narrative below to make the connection between the video they watched and their task explicit. A visualization of this staging webpage can be seen in Figure 1.

> You are on your way to a protest when you see on your social media feed the video from the previous screen, which appears to show a protestor being taken and driven away in an unmarked van. You recognize the video's background as close to the place where your protest is being held.
>
> You check the news, but it's not being covered by your local news station when you check.
>
> As you scroll further through your social media feed, you see that this video is the main topic of conversation.
>
> **Please consider which of the following posts in your feed you might think that other protestors should be aware of.**

After reading these instructions, the participant is prompted to click "Start," which reveals the simulated social media feed showing the coded replies, scraped from Twitter, discussing $R1$. As noted earlier there are four common rumour categories: affirmations, denials, neutral statements, and questions. Each post in the environment maps to one of these categories. Of the 86 replies for $R1$, a random sample of 10 affirming posts, 10 denying posts, 2 neutral posts and 2 questioning posts are shown to participants. Each post also contains a reshare button in the bottom right which the participant can choose to click on, indicating their intent to reshare that information. Posts are assigned a random profile image (male vs. female presenting) and name, using the Random User Generator[2] API, as well as a random timestamp which they are ordered by. Relative to previous survey-based approaches, the mock environment is meant to closely mimic the look and functionality of a typical social media platform, improving ecological validity. Participants

[2]https://randomuser.me/

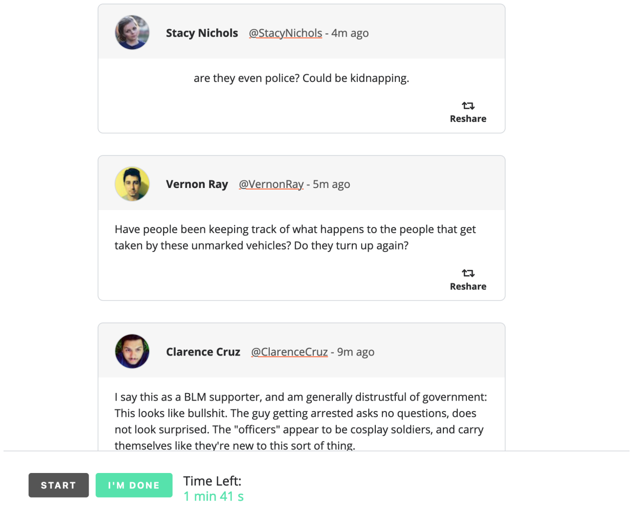

Figure 2: Mock environment showing posts discussing $R1$. The profile image, name assigned to the user posting, and timestamp are all randomly generated.

assigned to the treatment group will have three-quarters of the posts that deny $R1$ tagged with a credibility indicator. These posts misleadingly suggested that the arrest in the video was staged.

Additionally, given the raucous and hectic nature that can define protests, it may be the case that demonstrators do not have a sufficient amount of time to properly assess all the information that they are exposed to (Huang et al., 2015). To account for this threat to ecological validity, participants have a two minute time limit imposed on them to reshare posts. A screenshot of the mock environment can be seen in Figure 2. Once time has run out or the participant indicates that they're finished resharing posts, they will be thanked and taken to the next scenario.

Participants will complete a similar procedure for $R2$. They will watch the original video that catalyzed the contact tracing rumour, answer an attention check and whether they had seen the video before, and then be given their second narrative for the second protest scenario.

> On your way home, you get a notification that your local government is holding a press conference to discuss the protest you were just at.
>
> You tune in and see an official casually mention that they are "contact tracing" arrestees and seeing if they have links to organized crime.
>
> Most of your social media feed is discussing what this press conference means for tomorrow's protest, which you are planning on attending.
>
> **Please consider which of the following posts in your feed you might think that other protestors should be aware of.**

After doing so, they will again be able to press "Start" and be given two minutes to share posts pertaining to $R2$. Similarly, of the 81 replies scraped from Twitter discussing $R2$, a random sample of 10 affirming posts, 10 denying posts, 2 neutral posts, and 2 questioning posts are used.

Once time has run out or the participant indicates that they're finished resharing posts, they will be thanked and taken to the post-study questionnaire. There will be two questions where they are asked to estimate the likelihood of each rumour being true. This is to measure the effect of the treatment on participants' perception of the rumours' veracity. They are also asked how they decided which posts to reshare and, if they were assigned to the treatment group, how the credibility indicator factored in to their decision making process. Finally participants are asked if they had any issues completing the study, if they consent to their answers being released in an anonymized dataset, and if they had any final comments. Once completed, participants are redirected back to the recruitment platform to collect their reward.

Credibility Indicator Intervention. While this study attempts to retain a conceptual distinction between rumour and misinformation through a Bayesian framework, the relevant literature on misinformation's correction can provide the foundation for a targeted credibility indicator. In the context of rumour systems, it is rarely the case that a rumour is exclusively affirmed or denied online. Thus, in accordance with Lewandowsky et al. (2012) these contrasting narratives can serve as the basis for the credibility indicator, with additional content that fosters a healthy skepticism of information. Therefore, a credibility indicator directly on misleading posts that cites contrasting, alternative and local narratives, while noting that it is hard to verify information and easy to misinterpret events during protests, addresses key cognitive biases.

Much of the relevant literature use credibility indicators that cite professional fact-checking organizations as the warning's rationale (Yaqub et al., 2020; Friggeri et al., 2014; Pennycook, McPhetres, et al., 2020). It is unlikely that experts would be able to evaluate the relative evidence levels of rumours during ongoing protests, or determine who the experts ought to be; however, preliminary research has shown that crowdsourcing the truthfulness of propositions can be done with high levels of accuracy (Micallef et al., 2020). In addition, a recent study by Yaqub et al. (2020) has shown that individuals respond in roughly equal measures to professional fact-checks as crowdsourced ones. This form of credibility indicator is currently being piloted by the social media platform Twitter to stem misinformation. While research needs to be done to see if crowdsourced fact-checks can generalize to esti-

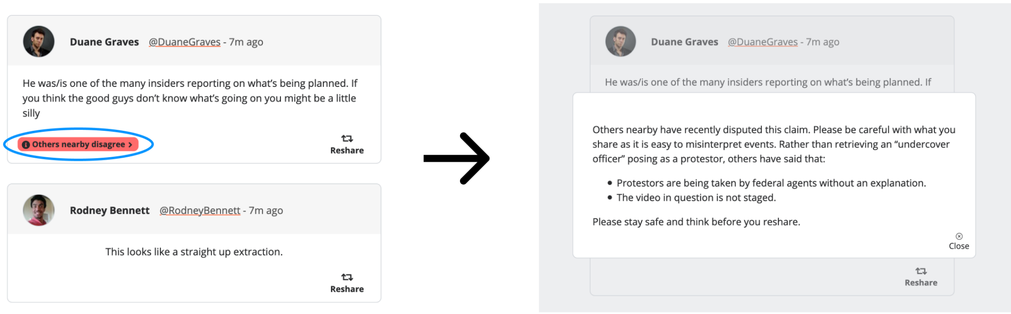

Figure 3: Contrastive intervention which suggests alternative, local narratives to misleading content.

mating evidence levels in rumour systems, it provides a useful basis for this study's intervention.

To summarize, this study's intervention, finetuned to the context of protest rumour systems, will be a brief, contrastive, crowdsourced credibility indicator suggesting that others disagree with that post's statement. Upon clicking the credibility indicator, a modal window will appear linking to contrastive narratives surrounding the rumour. The modal window approach to credibility indicators has been widely implemented by social media platforms and allows for an unintrusive tag that, when clicked on, provides greater context to the credibility indicator's rationale. An example of the proposed credibility indicator is shown in Figure 3.

4.5 Participants

Participant recruitment was done via Prolific. Prolific allows researchers to recruit participants based on various inclusion criteria. American and Canadian participants who use social media were recruited and paid £8.20 per hour for an estimated 10 minute experiment. After completing the experiment, the exclusion criteria are participants who: do not use social media, were previously aware of either rumour, fail either of the attention checks or who are statistical outliers in completion time. Attending protests was originally a criterion for inclusion; however, the importance of allowing participants to not reveal such sensitive information led to dropping this criterion. As Prolific only provides rudimentary demographic data along with a unique identifier for each participant, no personally identifiable information was collected. After verifying each participant's results and compensating them, participants' Prolific IDs were encrypted and the original IDs removed from the dataset.

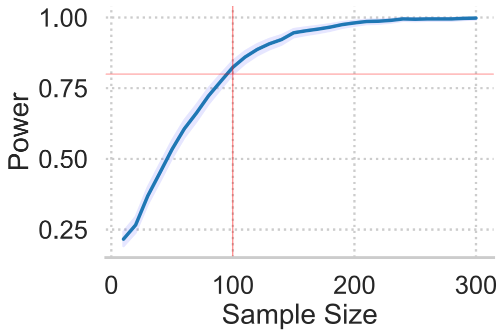

Figure 4: Power simulation results focusing on H_1 and H_3, regressing on p_R^a. Horizontal red line represents the power threshold ($\beta = 0.8$). Vertical red line represents the minimum sample size to pass the power threshold. At each sample size 1000 trials were conducted.

Power Analysis. To ensure that the study has sufficient power, a Monte-Carlo simulation was run using mock-data that affirms H_1 and H_3. This was done by training a linear mixed-effects model with p_R^a as the dependent variable, and the condition assignment, evidence level and their interaction as fixed effects. The participant ID was incorporated as a random effect to account for repeated measures for high and low evidence rumours (Brauer & Curtin, 2018; Yaqub et al., 2020). The coefficient of interest, and effect size, is the interaction term and was set to be -0.2 — indicating a one fifth reduction in the affirmation of low-evidence rumours. In order to achieve a power of 0.80 ($\alpha = 0.05$) the study must recruit at least 100 participants. To account for exclusion criteria and participant drop out, and in accordance with budgetary constraints, 189 participants were recruited. The results from the power simulation, which was conducted with SIMR (Green & MacLeod, 2016) are visualized in Figure 4.

5 Results

After removing participants based on the exclusion criteria, the final dataset contains 107 participants, each exposed to 56 posts, totaling 5,992 "resharing decisions." An exclusion criterion that was omitted in the preregistration, but deemed relevant, is to remove participants who did not reshare any posts. These participants indicated that they act as observers on social media and do not share content. Thus, they are unlikely to contribute to the spread of misleading information. The study ran from June 3, 2021 until June 7, 2021. As indicated in Table 2, the results from Prolific indicate a diverse range of participants across age, gender, political orientation, and affiliated social movements. Participants were distributed nearly uniformly

		Missing	Overall	Condition: Control	Condition: Treatment	p value
N			107	54	53	
Age		0	31.6 (12.2)	30.5 (11.7)	32.7 (12.6)	0.33
Education	High school	0	23	11	12	0.57
	Community college		15	6	9	
	Undergraduate degree		51	30	21	
	Graduate degree		15	6	9	
	Doctorate degree		2	1	1	
	None		1	0	1	
Political affiliation	Left	0	30	16	14	0.17
	Centre-left		35	23	12	
	Centre		14	4	10	
	Centre-right		11	5	6	
	Right		5	2	3	
	None		12	4	8	
Attends protests?	No	7	82	41	41	0.67
	Yes		18	8	10	
Gender	Woman	0	56	29	27	0.82
	Man		48	24	24	
	Non-binary		3	1	2	
Affiliated movements	Racial equity	0	91	48	43	0.99
	Climate change		82	44	38	
	Gender equity		81	41	40	
	Free speech		70	36	34	
	LGBTQA+ rights		65	35	30	
	Indigenous rights		59	33	26	
	Labour rights		57	27	30	
	Religious rights		43	23	20	
	Small government		23	13	10	
	Other		3	2	1	
	None		2	0	2	
Social medias	Instagram	0	87	47	40	0.72
	Facebook		82	41	41	
	Reddit		76	39	37	
	Twitter		59	29	30	
	Tiktok		42	26	16	
	Snapchat		40	22	18	
	WhatsApp		1	0	1	
	Clubhouse		1	1	0	
	Telegram		1	0	1	
	Tumblr		1	1	0	

Table 2: Descriptive statistics for participants, segmented by condition assignment. Mean age is reported along with standard deviation in parentheses. p values are calculated using a two sample T-test for continuous variables and chi-square test for categorical variables.

between the treatment ($N = 54$) and control ($N = 53$) conditions. Additionally, no participants indicated significant trouble completing the study, and only three participants failed an attention check — indicating a high quality of data.

Of the 107 participants, only 18 indicated that they had attended a protest in the past two years. As a result, to avoid reducing statistical power non-protestors were included in the analyses. The anonymized data used in these analyses, less those who did not want their responses shared, is available online (see, https://osf.io/6ug4m/).

First, this section will present the confirmatory analyses, which test hypotheses H_1–H_4 based on Krafft and Spiro's rumour proportions framework. After conducting these *system-level* analyses, exploratory analyses are conducted at the *post-level*. These exploratory analyses investigate how findings from previous studies map to rumour systems and provide the basis for future research.

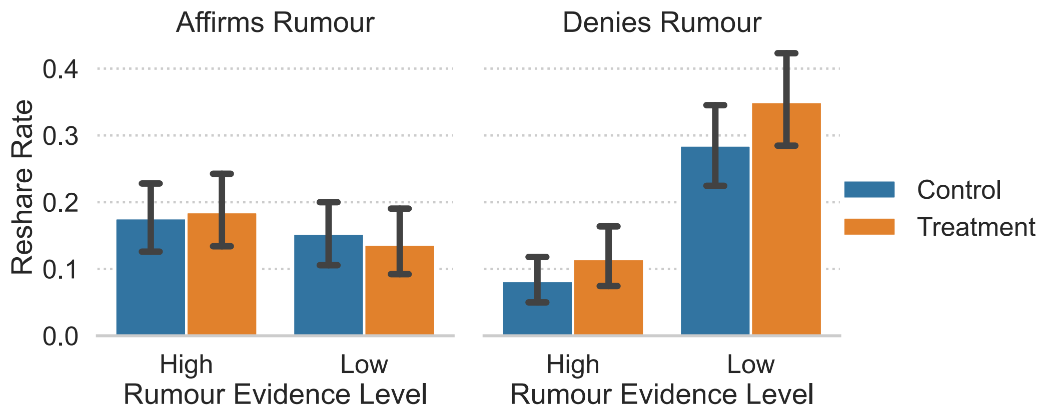

Figure 5: Percentage of posts shared affirming and denying, high and low evidence rumours for each participant. Error bars show 95% confidence interval.

5.1 Confirmatory Analyses

Of interest for the preregistered hypotheses (H_1–H_4) are the affirmation rate (p^a_R) and denial rate (p^d_R) for the high evidence rumour ($R1$) and low evidence rumour ($R2$). For all participants, p^a_{R1}, p^a_{R2}, p^d_{R1} and p^d_{R2} are aggregated and visualized in Figure 5. The left plot visualizes p^a_R for the high ($M_{(c)ontrol}$=0.18, SD_c= 0.19; $M_{(t)reatment}$=0.19, SD_t=0.20) and low (M_c=0.15, SD_c=0.19; M_t=0.14, SD_t= 0.18) evidence rumours, and the right plot visualizes p^d_R for the high (M_c=0.08, SD_c= 0.12; M_t=0.11, SD_t=0.15) and low (M_c=0.28, SD_c=0.22; M_t=0.35, SD_t=0.16) evidence rumours. These statistics are segmented by condition assignment.

Contrary to what was observed in the coded Twitter discussions, this indicates a system in proportion according to Krafft and Spiro, as $p^a_{R1} > p^a_{R2}$ and $p^d_{R1} < p^d_{R2}$. However, many participants exclusively reshared posts affirming $R2$, or denying $R1$, indicating that interventions may still be useful.

Model Estimation. To formally evaluate H_1 and H_3, a linear mixed-effects model was trained with p^a_R as the dependent variable, and the condition assignment, evidence level and their interaction as fixed effects. As mentioned earlier, there are two measures of p^a_R for each participant, p^a_{R1} and p^a_{R2}. To account for repeated measures, the participant ID was included as a random effect (Yaqub et al., 2020; Brauer & Curtin, 2018). H_2 and H_4 were similarly assessed using p^d_R as the dependent variable. While the preregistration indicated that demographic controls may be included as parameters, none were significant or meaningfully altered the treatment effect estimate and thus were omitted for clarity in the final regressions. Both models were fit using restricted maximum likelihood (REML) and all p values were estimated via t-tests using the Satterthwaite approximations to degrees of freedom. To account

	Dependent variable:	
	Affirmation Rate (p_R^a)	Denial Rate (p_R^d)
	(1)	(2)
Constant	0.324***	0.185***
	(0.037)	(0.032)
(A) Low Evidence	−0.034	0.301***
	(0.038)	(0.044)
(B) Treatment	0.022	0.052
	(0.054)	(0.047)
(A) X (B)	−0.056	0.008
	(0.055)	(0.062)
Observations	202	202
Log Likelihood	-15.080	2.994
Akaike Inf. Crit.	42.160	6.012
Bayesian Inf. Crit.	62.010	25.862
		*p<0.1; **p<0.05; ***p<0.01

Table 3: Linear mixed-effects model demonstrating the effect of treatment on the affirmation and denial rate of rumours. Parameter standard errors are reported in parentheses.

for positive skew in both dependent variables, both p_R^a and p_R^d were square-root transformed. This ensured that the distribution of residual terms followed a normal distribution. As with the dataset used, the scripts used to estimate these models are available on the study's preregistration.

Regression summary. H_1 and H_3 predicted that the treatment would decrease p_R^a for low evidence rumours and increase p_R^a for high evidence rumours respectively. By investigating the coefficient estimating the interaction between the treatment condition and the rumour evidence level, it is clear that this was not supported ($\beta = -0.06, SE = 0.06, p = 0.33$).

Similarly, H_2 and H_4 predicted that the treatment would increase p_R^d for low evidence rumours and decrease p_R^d for high evidence rumours respectively. This also was not supported ($\beta = 0.01, SE = 0.06, p = 0.89$). Thus, it is clear that the intervention did not meaningfully alter the sharing of misleading information in the mock-protest scenario. The full table presenting the regression results can be seen

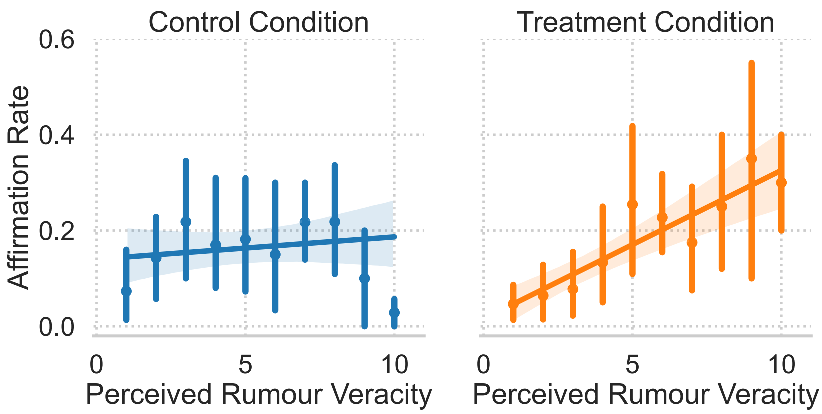

Figure 6: Scatterplot (with best-fitting regression line) showing the association between perceived rumour veracity and p_R^a. The error band on the regression represents the 95% confidence interval. For clarity, each scatter point represents the mean affirmation rate for that discrete value along the x-axis, with a 95% confidence interval.

in Table 3.

Attention is all you need? As previously noted, cognitive-inattention based interventions suggest that increasing the salience of veracity — relative to other factors like novelty or political concordance — is a key mechanism by which credibility indicators are effective. Thus, one explanation for why this study's intervention was not effective is that the treatment failed to increase the salience of veracity in the resharing calculus. Using the participant's veracity estimates from the post-study questionnaire, I can investigate this possibility. This is measured, for each rumour, on a scale from 1–10, where a value of 1 indicates the participant estimating the rumour to be "completely false" and 10 indicates estimating the rumour to be "completely true." Figure 6 visualizes the relationship between the estimated veracity and p_R^a for each participant and rumour, segmented by condition assignment. As can be seen visually, there is a much stronger association between perceived veracity and sharing behaviour in the treatment group.

To formalize this relationship, a similar linear mixed-effects model was trained, using the square-root transformed p_R^a as the dependent variable; the condition assignment, the perceived rumour veracity, their interaction, and the rumour evidence level as fixed effects; and the participant ID as a random effect. As can be seen in the regression summary in Table 4, there is a significant positive interaction between the condition assignment and the perceived rumour veracity on p_R^a ($\beta = 0.03, SE = 0.01, p < 0.01$). These results indicate that the intervention made participants more likely to share in accordance with their perceived veracity judge-

	Dependent variable:
	Affirmation Rate (p^a_R)
Constant	0.287***
	(0.053)
Low Evidence	−0.070***
	(0.026)
(B) Treatment	−0.151**
	(0.074)
(C) Perceived Rumour Veracity	0.011
	(0.008)
(B) X (C)	0.032***
	(0.012)
Observations	199
Log Likelihood	-8.799
Akaike Inf. Crit.	31.598
Bayesian Inf. Crit.	54.651

*p<0.1; **p<0.05; ***p<0.01

Table 4: Linear mixed-effects model demonstrating the effect of treatment and perceived rumour veracity on a rumour's affirmation rate. Parameter standard errors are reported in parentheses.

ments. Thus, in line with Pennycook et al. the intervention increased the salience of veracity in the sharing calculus (2021). However, it also implies that a cognitive-inattention based theory is insufficient to explain the null results.

Initially, it may appear incongruent that the introduction of credibility indicators would make participants share in greater accordance with their veracity estimates, but fail to alter sharing behaviour. However, these findings can be reconciled *if the intervention does not alter the participants' veracity estimates.* Indeed, when training a final linear mixed-effects model using the participant veracity estimate as the dependent variable; the evidence level, condition assignment, and their interaction as fixed effects; and the participant ID as a random effect, we see that there is no effect of the treatment on how plausible the participants perceived each rumour ($\beta = -0.68, SE = 0.68, p = 0.32$). See Appendix B for the full regression summary. This relationship is visualized in Figure 7.

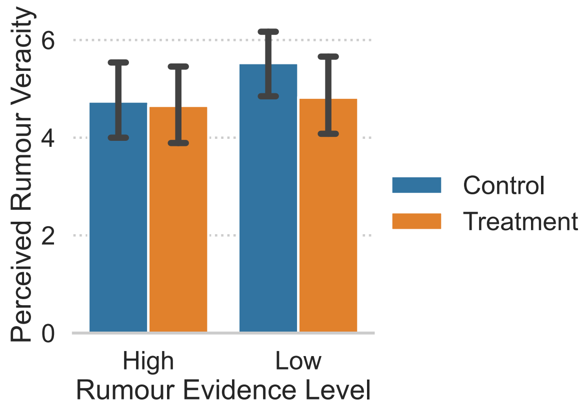

Figure 7: Participant veracity estimates for high and low evidence rumours. Error bars show 95% confidence interval.

Summary. This section presented findings from a preregistered experiment exploring the impact of credibility indicators on sharing behaviour during a mock-protest scenario. Of interest were system-level metrics outlined by Krafft and Spiro, which capture the rate at which propositions are affirmed or denied relative to their plausibility. In this study, there is weak evidence to suggest that the introduction of credibility indicators had a meaningful effect on the system-level metrics of interest. Importantly, this is not because participants failed to consider what was accurate when resharing, as the treatment increased the relationship between perceived veracity and rumour affirmation. Instead, a more plausible explanation is that participants were sharing *what they considered to be accurate*, but their veracity estimates remained biased despite the treatment. This raises theoretical and practical implications for the prospect of cognitive-inattention based treatments, which are considered in the discussion section.

5.2 Exploratory Analyses

Finally, I present exploratory analyses, at the post-level, to investigate three research questions. Given that credibility indicators are attached to individual posts, such analyses are especially pertinent. Of note, only 22.22% of participants in the treatment condition clicked on a credibility indicator to view the alternative, contrasting narrative to misleading content. In total, credibility indicators were only clicked on 3.30% of the time.

The first two research questions ask how post attributes, like the post source (RQ_1) or temporal proximity (RQ_2), affect sharing behaviour during protest demonstrations. The third research question (RQ_3) investigates any evidence of an implied truth effect, whereby attaching credibility indicators to a *subset* of false posts is considered to imply truthfulness in the posts without a credibility indicator.

Rather than being preregistered, confirmatory hypotheses, these analyses serve to provoke future research paths at the intersection of human-computer interaction and information studies.

Post attributes. Given the constraints imposed by the experimental design, there are two post attributes that were randomized and thus can be causally analyzed with respect to sharing behaviour: the post source and temporal proximity. These were both analyzed by Huang et al. (2015) in the context of the Boston Marathon Bombing and subsequent social media discussions. A key finding of this research was that in the absence of established sources, individuals trusted those in their personal network when broadcasting updates online. Thus, they found that where a piece of information originated from in one's friend network had a qualitative impact on perceived veracity and sharing behaviour. Due to privacy concerns, the mock-environment participants interacted with in this study was not connected to their personal social media accounts, and thus I am unable to make use of these friend networks. However, it is possible that other source attributes, like the gendered profile image (see, Figure 2) assigned to each post, could impact sharing behaviour and will be analyzed in this context.

Additionally, Huang et al. found that temporal proximity was associated with rumouring. This aligns with Earl et al.'s work, which found that demonstrators relied on temporally relevant information, broadcast on social media in real time. We may expect, then, that posts randomly assigned to be "more recent" via timestamps to be perceived as more relevant, and thus more likely to be shared. This is informally demonstrated in Figure 8, which visualizes the reshare rate for posts as a function of their assigned timestamp. While participants were not questioned on specific content in the post-study questionnaire, their sharing patterns can give insights and prompts for future research.

Implied truth effect. Finally, this exploratory analysis investigates the possibility of an implied truth effect, whereby attaching credibility indicators to a subset of posts is considered to imply truthfulness in the posts without a credibility indicator. As noted by Pennycook, Bear, et al. (2020), it is much easier for misleading information to be propagated than verified. Therefore, while a community-driven approach to rumour management, where individuals close to the rumour's origin provide evidence for and against propositions, is preferable — the nature of rumours imply a level of ambiguity in attaching credibility indicators to social media posts. Thus, even with a robust, community-driven approach to rumour management, it is still

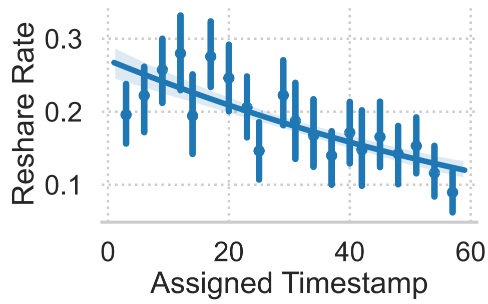

Figure 8: Scatterplot (with line representing log odds ratio) showing the association between temporal proximity and reshare rate. The error band on the regression represents the 95% confidence interval. For clarity, each scatter point represents the mean reshare rate for each 3 minute increment along the x-axis, with a 95% confidence interval.

likely that harmful content will go untagged. If there is an implied truth effect then the benefits of credibility indicators could be negated, as individuals may simply shift to sharing misleading posts that fail to be tagged. This is especially relevant given the short life cycle with which information is needed during protests. As currently implemented on social media platforms, the labelling of misinformation can take 10-20 hours (Shao, Ciampaglia, Flammini, & Menczer, 2016). Thus, content may proliferate before it can be tagged. Given that only three quarters of misleading posts (*i.e.*, affirming a low evidence rumour or denying a high evidence rumour) were tagged with credibility indicators, I am able to test whether the untagged posts fell prey to the implied truth effect.

Model estimation. The dependent variable in these analyses is the decision to reshare a given post. Since the dependent variable is a binary choice between the choice to reshare or not reshare, I employed a binomial logistic regression (BLR). To account for repeated measures for varying items (56 observations per participant), the participant ID and post ID, derived from the Twitter API, were incorporated as random effects. Whether the post is misleading, defined as affirming the low evidence rumour or denying the high evidence rumour, the source gender (male vs. female profile image), and timestamp randomly assigned to each post are included as fixed effects. Similar to Pennycook, Bear, et al. (2020), the control condition is used as the baseline and tests for a warning effect with a "warned" dummy that indicates the post being in the credibility indicator treatment and having a warning. Finally, I test for an implied truth effect with an "untagged" dummy that indicates a post being in the credibility indicator treatment and not having a warning. An

	Dependent variable:
	Reshared?
Constant	-1.131^{***}
	(0.195)
Warned	-0.141
	(0.255)
(D) Misleading	-0.784^{***}
	(0.188)
(E) Untagged	0.134
	(0.215)
(D) X (E)	0.453^{*}
	(0.249)
Male	0.011
	(0.076)
Timestamp	-0.023^{***}
	(0.002)
Observations	5,992
Log Likelihood	-2,416.299
Akaike Inf. Crit.	4,850.599
Bayesian Inf. Crit.	4,910.882
	*p<0.1; **p<0.05; ***p<0.01

Table 5: BLR mixed-effects model demonstrating the effect of post attributes and the implied truth effect on resharing odds. Parameter standard errors are reported in parentheses.

interaction term between the untagged dummy and whether the post is misleading is also included. As with the dataset, the script used to train the BLR is available on the study's preregistration.

Regression summary. RQ_1 asks how source attributes affect resharing behaviour. Given constraints imposed by the experimental design, the focus then is on the male or female presenting profile image randomly assigned to each post. As can be seen in Table 5, there was no effect of the source gender on resharing odds $(\beta = 0.01, SE = 0.08, p = 0.88)$.

Another post attribute, which was expected to be associated with rumouring, was the assigned timestamp, indicating temporal proximity. Confirming what was visualized in Figure 8, there was a statistically significant effect of temporal proximity on resharing odds, with less recent posts being reshared at a lower rate ($\beta = -0.02, SE < 0.01, p < 0.01$).

Finally, RQ_3 posited the possibility of an implied truth effect. By investigating the coefficient for the "untagged" dummy one can see that there is no main effect ($\beta = 0.13, SE = 0.22, p = 0.53$). However, there is a positive interaction effect between a post being untagged and that post being misleading ($\beta = 0.45, SE = 0.25, p = 0.07$). Misleading posts that failed to be tagged were reshared at a higher rate than those that were — though both were reshared at a lower rate than posts that were not misleading. Thus, there is weak evidence to support an implied truth effect.

Summary. This section proposed and investigated three exploratory analyses. In doing so, the effect of source gender, temporal proximity, and failing to tag all misleading posts were probed. In this study, the assigned profile image, presenting as either male or female, had little effect on resharing behaviour. However, temporal proximity was found to be significantly associated with rumouring, and mixed evidence was found to support an implied truth effect. Both findings have practical and scholarly implications, which are discussed in tandem in the discussion section.

6 Discussion

This preregistered study set out to investigate how design factors, like the addition of credibility indicators, alters sharing behaviour in rumour systems. The case of interest for this study being protest demonstrations in a North American context. In doing so, I examined how — given rumours with varying degrees of evidence — credibility indicator interventions affect key system-level metrics. In addition, I presented exploratory analyses which investigated the effects of source attributes and an implied truth effect. These analyses leverage a minimal clone of a Twitter-like platform I implemented from scratch for the purposes of conducting behavioural experiments, increasing ecological validity. Through fine-tuned control of the experimental environment, real-world stimuli scraped from Twitter, and carefully chosen attention checks I was able to collect high quality behavioural data for analysis. When situated within a regression framework, the results present a nuanced picture with respect to the efficacy of credibility indicator interventions in rumour systems.

The confirmatory analyses focused on the rate at which participants affirm or deny rumours in a protest scenario. The results show that credibility indicator interventions strengthened the relationship between perceived veracity and sharing behaviour. Thus, participants in the treatment were more likely to take into consideration how accurate they thought each post was when deciding what to reshare. Simultaneously, the credibility indicators failed to meaningfully alter sharing behaviour. It was then argued that this is likely a result of the intervention failing to alter participants' veracity estimates.

The exploratory analyses then moved beyond aggregate, system-level metrics to analyze the effect of credibility indicators at the level of the post. These preliminary findings uncovered a significant association between temporal proximity and resharing. Conversely, other source attributes, like the gender associated with each profile image, had a negligible impact on sharing behaviour. Finally, these exploratory analyses found preliminary evidence for an implied truth effect in rumour systems.

Each of these findings have methodological, theoretical, and practical implications. First, this thesis' contributions to the field of human-computer interaction and study of misinformation are outlined. Then this study's limitations are discussed. In each subsection, prompts for future research are presented which would bolster the robustness of this study.

6.1 Contributions

This paper developed three key contributions to the study of rumour, misinformation and human-computer interaction — each with methodological, theoretical, and practical implications. First, this study moved beyond a strict true or false binary that is likely problematic. Second, this study's null results help delineate the settings conducive to cognitive inattention-based interventions. And finally, this study provides preliminary experimental evidence that demonstrates the importance of temporal proximity in rumouring.

Contribution one: Moving beyond a true/false binary. This study adopts a Bayesian framework for information management, first proposed by Krafft and Spiro. This approach contrasts the dominant paradigm of misinformation research, which focuses on verifiably true or false information (*e.g.*, "Pope Francis shocks world, endorses Donald Trump for president"). This differs from unverified information but the two are often conflated, or unverified information is ignored entirely. Rather than stifling all unverified information, this study's rumour proportions approach adds much-needed nuance to the study of misinformation to include the superset of all unverified information, of which proven true or false claims are a special case. Thus, a key contribution of this study is operationalizing a rumour proportions framework in an experimental setting.

However, despite this incremental improvement, this study only made use of rumours that have been evaluated retroactively. In temporally constrained contexts, like protest demonstrations, relying on third-party organizations to provide evidence for or against propositions would neither be feasible or desirable. From a practical viewpoint, this study proposed a crowdsourced mechanism for aggregating evidence estimates. This could be implemented as a voting feature on social media, where community members and demonstrators are able to provide evidence for and against emerging narratives. As mentioned earlier, this is being piloted by the social media platform Twitter, and crowdsourcing accuracy judgements have been shown to be unbiased in various contexts (Pennycook & Rand, 2019; Allen et al., 2020; Micallef et al., 2020). However, future research still needs to be done on how such approaches would translate to rumour systems. Despite this, moving beyond a true/false dichotomy represents a key incremental contribution of this study.

Contribution two: The limits of increased attention. The main finding of this study is that credibility indicators made participants share in greater alignment with what they perceived as veracious but did not alter those perceptions themselves. Thus, those with biased veracity estimates of the two rumours still shared misleading information. This resulted in null findings, where the treatment did not alter the system-level metrics of interest. In another light, the degree to which resharing habits reflect known evidence can be viewed as a function of the accuracy of one's veracity judgements and the degree to which veracity factors into one's decision to reshare information, relative to other factors like novelty or political concordance. It appears the interventions from this study had a meaningful effect on the latter — while having weak evidence, or possibly a much smaller effect, on the former.

This finding poses a complication to Pennycook et al.'s cognitive-inattention based interventions — which presuppose that individuals have access to reasonable inferences based on evidence, but are not taking them into account. However, for a variety of reasons such inferences may be biased. In these cases, increasing the saliency of veracity is unlikely to be effective. This is likely to be the case in rumour systems, where events can unfold, and narratives can be distorted, very quickly — but also in cases where specialized knowledge is required to evaluate information, as with scientific or medical misinformation. In these cases, the intervention must also encourage deliberation and a healthy skepticism as to why some piece of information may be misleading, rather than simply increasing the significance of prior veracity judgements. Thus, a key contribution of this study is to delineate the limits of addressing cognitive-inattention in stemming the spread of misleading information.

A key question for future research is to uncover why credibility indicators made participants share in greater alignment with their veracity perceptions, but did not alter those perceptions themselves. I propose three preliminary hypotheses for why this was the case. First, it could be the case that it is simply easier to address cognitive-inattention than it is to change ones view of what is true. This would explain why there is stronger evidence for the increased salience of veracity and weak evidence for the effect on sharing patterns. Second, the null-results could be a function of the crowdsourced framing of the credibility indicators. In the open-ended questionnaire, many participants noted that they were not swayed by the fact that others nearby disagreed with a piece of information. One said, "I think for myself, so 'others nearby disagree' is meaningless to me." In this scenario, the credibility indicator reminded participants that sharing accurate information is important, but did not, in their mind, provide sufficient evidence to alter their view of what was true. Finally, an extremely small portion of those in the treatment group ever clicked on the credibility indicator, and thus only received a part of the message. Not viewing the modal and understanding the rationale behind the credibility indicator could limit the effectiveness of the treatment. This is of practical and scholarly importance as credibility-indicators are often implemented as an unobtrusive tag that links to a modal or third-party website which provides more comprehensive information. Thus, future research may evaluate how other forms of interventions — which make the contrastive, counter-narrative more apparent could aide in this respect.

Contribution three: The importance of temporal proximity. A final key contribution that this study makes to the field of human-computer interaction and misinformation relates to the role of temporal proximity and rumouring online. Similar to Huang et al., this thesis outlined empirical evidence that temporal proximity is associated with resharing behaviour. This is of practical import in rumour systems as information is often needed within a short timeframe to be useful (Earl et al., 2013). This is especially relevant given the possibility of an implied truth effect. Thus, the ability to quickly evaluate which posts are misleading and which are not is of utmost importance. Given the lag that exists between post creation and tagging, attaching credibility indicators to posts directly may be infeasible (Shao et al., 2016). If this is the case then system-wide interventions (*e.g.*, a banner at the top of the social media page asking individuals to post deliberately) may be more practically feasible and relevant.

6.2 Limitations

While demonstrating meaningful scholarly contributions, this thesis has limitations. First, when conducting the power analysis the intended effect size was -0.2, indicating a one fifth reduction in the affirmation of low-evidence rumours. While selected to represent a meaningful reduction in the spread of misleading information, upon completion of this study it is clear that such a large effect was optimistic. Thus, it is likely that this study is under-powered. Future work can compensate for this by estimating smaller effect-sizes and increasing the power of subsequent experiments. In addition, this study only focused on two rumours. To reduce rumour specific variation, further work ought to expand the set of propositions analyzed.

Third, this study only considers information posted in good faith. As with all aspects of social media, coordinated attacks are likely and could undermine public trust in such interventions. This poses practical challenges to the implementation of credibility indicators, as malicious actors could have unverified information flagged unduly. Similarly, the event of incorrectly flagging information as misleading, even if unintentionally, is a real possibility. Inappropriate credibility indicators may undermine their efficacy more broadly. The effect of incorrectly tagging information as misleading on subsequent sharing behaviour was not observed in this study, but is relevant as this is a likely outcome of implementing credibility indicators in rumour systems.

Finally, in environments where official sources are unavailable, individuals will often turn to the user posting the information as an indicator of credibility. For example, people are likely to trust information shared by those they know personally (Huang et al., 2015). Additional research on political concordance and misinformation suggests that those would be more likely to trust information shared by those who signal an aligned political affiliation (Mosleh et al., 2021). A future extension of this work could be to manipulate the profiles assigned to social media posts in a way that includes cues as to the source credibility; for example, assigning profiles left-indicating or right-indicating usernames. Another extension that incorporates source credibility would be to connect the experimental environment with the participant's social media account and assign profiles based on their real-world connections. Of course, making clear that any the participants' responses are private and that posts "shared" in the experimental environment are not actually shared to their friend network would be incredibly important. To maintain a clear conceptual distinction between the mock-environment and participants' own social media accounts, this study did not attempt such integrations.

7 Conclusion

Each year, millions flood the streets and attend protests, exacting a core promise of democracy. To do so, they use social media and direct messaging as foundational infrastructure to organize, broadcast updates, and alert others to potential dangers. However, the harms that can result from the spread of misleading information in, what are often, precarious situations can be equally grave. This tension is compounded by the ambiguity of protests, as information is often impossible to verify. Despite this, little experimental research has been done that asks how the structure and design of online platforms affect the information that demonstrators spread and have access to. Interventions that improve the quality of information spread during protests could have tangible benefits for demonstrators and community members.

This preregistered study examined the impact of credibility indicators on the spread of two protest rumours. In doing so, it sought to evaluate how such design interventions affect the rate at which rumours are affirmed and denied, relative to the rumours' evidence levels. This study then presented exploratory analyses, at the post-level, to analyze how source attributes impacted resharing behaviour, and the possibility of an implied truth effect. To answer these, 189 participants were recruited to complete a behavioural experiment simulating a mock-protest environment, which was built from scratch. Participants were segmented into a control group and a treatment group, in which misleading posts were tagged with credibility indicators, and then presented with a high and low evidence rumour. The participant's task was to share information that they would feel is relevant given the protest scenarios.

Results evaluating the preregistered hypotheses suggest that the intervention was successful in focusing participants' attention on the veracity of information. However, the intervention failed to shift participants' perception of what was true. As a result, those in the treatment condition continued to share misleading information.

This study's exploratory analyses then demonstrated that, regardless of the actual content of a post, temporal proximity has an effect on resharing. This combined with evidence to support an implied truth effect, where the absence of a credibility indicator is assumed to imply truthfulness, means that misleading information may spread before it has a chance to be tagged, posing practical concerns.

From a policy perspective, this research suggests that encouraging the thoughtful deliberation of rumours requires more than focusing attention on accuracy. In situations where the veracity of information may be unknowable processes must be put in place that amplify the voices of those close to the information's source, inspire a healthy skepticism of information shared online, and encourage demonstrators to consider the impact of misleading information. Social media platforms are often designed to reduce friction between what one thinks and what one posts. However, in protest demonstrations, the speed of sharing must be balanced with the reflexivity needed for collective sensemaking.

Ultimately, further research is needed to understand how design factors affect knowledge integration and the spread of information. This ought to be done within specific contexts since, as evidenced by this study, different use cases dictate different practical and theoretical constraints. Through an interdisciplinary approach — melding community members, domain experts, researchers, and practitioners — a better information ecosystem can be cultivated.

References

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2020). Scaling up fact-checking using the wisdom of crowds. *Preprint available at https://psyarxiv.com/9qdza.*

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: Adapting the tse framework to big data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119.

Bordia, P., & DiFonzo, N. (2004). Problem solving in social interactions on the internet: Rumor as social cognition. *Social Psychology Quarterly*, *67*(1), 33–49.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389.

Brennen, J. S., Simon, F., Howard, P. N., & Nielsen, R. K. (2020). Types, sources, and claims of covid-19 misinformation. *Reuters Institute*, *7*, 3–1.

Buchanan, L., Bui, Q., & Patel, J. (2020, Jul). Black lives matter may be the largest movement in u.s. history. *The New York Times*. Retrieved from `https://www.nytimes.com/interactive/2020/07/03/us/george-floyd-protests-crowd-size.html`

Chandra, P., & Pal, J. (2019). Rumors and collective sensemaking: Managing ambiguity in an informal marketplace. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12).

Coppock, A. (2019). Generalizing from survey experiments conducted on mechanical turk: A replication approach. *Political Science Research and Methods*, *7*(3), 613–628.

De Choudhury, M., Jhaver, S., Sugar, B., & Weber, I. (2016). Social media participation in an activist movement for racial equality. In *Proceedings of the international aaai conference on web and social media* (Vol. 10).

Earl, J., McKee Hurwitz, H., Mejia Mesinas, A., Tolan, M., & Arlotti, A. (2013). This protest will be tweeted: Twitter and protest policing during the pittsburgh g20. *Information, communication & society*, *16*(4), 459–478.

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition*, *38*(8), 1087–1100.

Fine, G. A., & Turner, P. A. (2001). *Whispers on the color line: Rumor and race in america.* Univ of California Press.

Freelon, D., McIlwain, C., & Clark, M. (2018). Quantifying the power and consequences of social media protest. *New Media & Society*, *20*(3), 990–1011.

Friggeri, A., Adamic, L., Eckles, D., & Cheng, J. (2014). Rumor cascades. In *Proceedings of the international aaai conference on web and social media* (Vol. 8).

Gao, M., Xiao, Z., Karahalios, K., & Fu, W.-T. (2018). To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–16.

Garrett, R. K., & Weeks, B. E. (2013). The promise and peril of real-time corrections to political misperceptions. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 1047–1058).

Green, P., & MacLeod, C. J. (2016). Simr: an r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.

Huang, Y. L., Starbird, K., Orand, M., Stanek, S. A., & Pedersen, H. T. (2015). Connected through crisis: Emotional proximity and the spread of misinformation online. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing* (pp. 969–980).

Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, *10*(1), 1–20.

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

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of experimental psychology: Learning, memory, and cognition*, *20*(6), 1420.

Kelly Garrett, R. (2006). Protest in an information society: A review of literature on social movements and new icts. *Information, communication & society*, *9*(02), 202–224.

Krafft, P. M., & Spiro, E. S. (2019). Keeping rumors in proportion: managing uncertainty in rumor systems. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–11).

Lee, J. (2020). Were portland protesters 'kidnapped' by federal officers in unmarked vans? *Snopes*. Retrieved 2021-03-25, from `https://www.snopes.com/fact-check/feds-unmarked-vans-portland/`

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, *13*(3), 106–131.

Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th acm international on conference on information and knowledge management* (pp. 1751–1754).

Marwick, A., Kuo, R., Cameron, S. J., & Weigel, M. (2021). *Critical disinformation studies: A syllabus.* Center for Information, Technology, & Public Life (CITAP), University of North Carolina at Chapel Hill. Retrieved 2021-03-27, from `https://citap.unc.edu/critical-disinfo`

McCreadie, R., Macdonald, C., & Ounis, I. (2015). Crowdsourced rumour identification during emergencies. In *Proceedings of the 24th international conference on world wide web* (pp. 965–970).

Micallef, N., He, B., Kumar, S., Ahamad, M., & Memon, N. (2020). The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. *arXiv preprint arXiv:2011.05773*.

Morrison, S. (2020). Minnesota law enforcement isn't "contact tracing" protesters, despite an official's comment. *Vox*. Retrieved 2021-

03-25, from https://www.vox.com/recode/2020/6/1/21277393/minnesota-protesters-contact-tracing-covid-19

Mosleh, M., Martel, C., Eckles, D., & Rand, D. (2021). Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. In *Proceedings of the 2021 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3411764.3445642 doi: 10.1145/3411764.3445642

Nassauer, A. (2018). Situational dynamics and the emergence of violence in protests. *Psychology of violence*, *8*(3), 293.

Nekmat, E. (2020). Nudge effect of fact-check alerts: source influence and media skepticism on sharing of news misinformation in social media. *Social Media+ Society*, *6*(1), 2056305119897322.

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS quarterly*, 407–426.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, *66*(11), 4944–4957.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526.

Pew Research Center. (2020). *In views of u.s. democracy, widening partisan divides over freedom to peacefully protest.* Retrieved from https://www.pewresearch.org/politics/2020/09/02/in-views-of-u-s-democracy-widening-partisan-divides-over-freedom-to-peacefully-protest/ (Online; accessed 08 July 2021)

Poell, T., & Van Dijck, J. (2015). Social media and activist communication. *Poell, Thomas & José van Dijck (2015). Social Media and Activist Communication. In The Routledge Companion to Alternative and Community Media*, 527–537.

Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? a preregistered direct replication of pennycook et al. (2020). *Psychological Science*, *0*(0), 09567976211024535. Retrieved from https://doi.org/10.1177/09567976211024535 doi: 10.1177/09567976211024535

Rosnow, R. L. (1988). Rumor as communication: A contextualist approach. *Journal of Communication*, *38*(1), 12–28.

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A

platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web* (pp. 745–750).
Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(3), 1–42.
Shibutani, T. (1966). *Improvised news: A sociological study of rumor.* Ardent Media.
Sitaula, N., Mohan, C. K., Grygiel, J., Zhou, X., & Zafarani, R. (2020). Credibility-based fake news detection. In *Disinformation, misinformation, and fake news in social media* (pp. 163–182). Springer.
Starbird, K., Spiro, E., Edwards, I., Zhou, K., Maddock, J., & Narasimhan, S. (2016). Could this be true? i think so! expressed uncertainty in online rumoring. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 360–371).
Steinert-Threlkeld, Z. C., Mocanu, D., Vespignani, A., & Fowler, J. (2015). Online social networks and offline protest. *EPJ Data Science*, *4*(1), 1–9.
Trana, T., Valechaa, R., & Raoa, H. R. (2020). False claims hurt: Examining perceptions of misinformation harms during black lives matter movement. *Research Gate*. Retrieved from https://www.researchgate.net/profile/Thi-Tran-57/publication/352295604_False_Claims_Hurt_Examining_Perceptions_of_Misinformation_Harms_during_Black_Lives_Matter_Movement/links/60c247ad299bf1949f495c42/False-Claims-Hurt-Examining-Perceptions-of-Misinformation-Harms-during-Black-Lives-Matter-Movement.pdf
Van Laer, J., & Van Aelst, P. (2010). Internet and social movement action repertoires: Opportunities and limitations. *Information, Communication & Society*, *13*(8), 1146–1171.
Varol, O., Ferrara, E., Ogan, C. L., Menczer, F., & Flammini, A. (2014). Evolution of online user behavior during a social upheaval. In *Proceedings of the 2014 acm conference on web science* (pp. 81–90).
Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., … Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857).
Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N., & Patil, S. (2020). Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–14).
Zeng, L., Starbird, K., & Spiro, E. (2016). # unconfirmed: Classifying rumor stance in crisis-related social media messages. In *Proceedings of the international aaai conference on web and social media* (Vol. 10).

Appendices

Appendix A Twitter Rumour Codebook

Code	Details
Irrelevant	Discussion that is either not relevant to the rumour, or would not have meaningful context in the experimental environment. Example: "'Outside agitators' also known as the police," or "Use your words... Mommy."
Affirms	Affirming the rumour in question (*i.e.* it does look like federal agents are kidnapping/detaining protestors or it does sound like law enforcement are using contact tracing apps to track protestors). Example: "Are they even police? Could be kidnapping," or "When the true use of contact tracing is acknowledged and public health will never be trusted again. This is a major unforced error in language. Can't take this back."
Denies	Denies the rumour in question (*i.e.* federal agents are not kidnapping/detaining professors or law enforcement are not using contact tracing technology on protestors). Examples: "I'm calling it now... this SMELLS of PSYOP. The person being 'kidnapped' remains silent when asked to give... their name. This smells more of a STAGED extraction done purely for SHOCK VALUE. If this was a real protestor... why didn't he give his name on camera?", or "Could we not use the term 'contact tracing' to describe police investigations into criminal matters?"
Neutral	Does not make a claim on the truthfulness of the rumour. Example: "Surrounding the van. Block it's exit. If they injure somebody you document it and take your pick of top level attorneys ready to pounce on these morons. Feds don't drive minivans with golf hats on the dash. Ever heard of Miranda?"
Questions	Expresses uncertainty in the rumour. Example: "Is this what the 'contact tracing' that was meant for coronavirus is being used for? Or is this just mere coincidence of expression? Seems relevant. Please clarify."

Table 6: Codebook used when annotating scraped Twitter replies for protest rumours.

Appendix B Effect of treatment on veracity estimates

	Dependent variable:
	Perceived Rumour Veracity
Constant	4.742^{***}
	(0.385)
(A) Low Evidence	0.784^{*}
	(0.475)
(B) Treatment	-0.052
	(0.557)
(A) X (B)	-0.684
	(0.685)
Observations	199
Log Likelihood	-477.973
Akaike Inf. Crit.	967.946
Bayesian Inf. Crit.	987.706
	*p<0.1; **p<0.05; ***p<0.01

Table 7: Linear mixed-effects model demonstrating the effect of treatment and rumour evidence level on perceived rumour veracity. Parameter standard errors are reported in parentheses.