Uncovering the Spiral of Silence in online political deliberation

By

Gabriela Juncosa

Submitted to
Central European University
Department of Network and Data Science

In partial fulfillment of the requirements for the degree of Doctor of Philosophy in Network Science

Supervisor: Prof. Gerardo Iñiguez & Prof. Tiago P. Peixoto

Vienna, Austria
2024

Researcher declaration

I, Gabriela Juncosa certify that I am the author of the work *Uncovering the Spiral of Silence in online political deliberation*. I certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, the contributions of others. The thesis contains no materials accepted for any other degrees in any other institutions. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright.

Statement of inclusion of joint work

Chapter 2 is based on a research project in collaboration with Dr. Taha Yasseri, Dr. Júlia Koltai and Dr. Gerardo Iñiguez. Dr. Iñiguez and I conceived the project idea. I collected and pre-processed the data. All authors contributed to designing the data analysis strategy, which I implemented. Additionally, all authors participated in the interpretation and analysis of the results and the writing of the paper. Dr. Yasseri, Dr. Koltai, and Dr. Iñiguez endorse this statement with their signatures below.

Chapter 3 is based on a research project in collaboration with Ms. Saeedeh Mohammadi, Dr. Margaret Samahita and Dr. Taha Yasseri. Dr. Samahita and Dr. Yasseri conceived the project idea, designed the experimental setup, and provided the funding for the experiment. I was responsible for coding, running the experiment, and pre-processing the data. Dr. Yasseri and Ms. Mohammadi led the results analysis efforts. All authors contributed to the interpretation and analysis of the results. Dr. Yasseri, Ms. Mohammadi and I collaborated on writing the paper. Dr. Samahita, Dr. Yasseri, and Ms. Mohammadi endorse this statement with their signatures below.

Chapter 4 is based on a research project in collaboration with Dr. Angel Sánchez, Dr. Júlia Koltai, Dr. Taha Yasseri and Dr. Gerardo Iñiguez. Dr. Iñiguez and I conceived the project idea, while all authors contributed to designing the experimental setup. Dr. Iñiguez and Dr. Sánchez provided funding for the experiment. I was responsible for coding, running the experiment, and pre-processing the data. Dr. Koltai and I conceptualized the data analysis strategy, which I implemented. All authors contributed to the interpretation and analysis of the results and to the writing of the paper. Dr. Sánchez, Dr. Koltai, Dr. Yasseri, and Dr. Iñiguez endorse this statement with their signatures below.

Signature of Gabriela Juncosa, PhD Candidate:

Gabriela Juncosa

Date: Dec 9th, 2024

Signature of Dr. Gerardo Iñiguez, endorsing statement of joint work:

Date: Dec 4th, 2024

Signature of Dr. Júlia Koltai, endorsing statement of joint work:

Date: Dec 4th, 2024

Signature of Dr. Angel Sánchez, endorsing statement of joint work:

Date: Dec 4th, 2025

Signed by SANCHEZ
SANCHEZ ANGEL -
***4841** on date
04/12/2024 with a
certificate issued

Signature of Ms. Saeedeh Mohammadi, endorsing statement of joint work:

Saeedeh

Date: Dec 4th, 2025

Signature of Dr. Margaret Samahita, endorsing statement of joint work:

Margaret

Date: Dec 5th, 2025

Signature of Dr. Taha Yasseri, endorsing statement of joint work:

Taha Yasseri

Date: Dec 5th, 2025

Abstract

This thesis explores the intersection of digital media, social behavior, and political expression, focusing on the dynamics of opinion expression in online environments. The widespread use of social media platforms, such as YouTube, has transformed public communication by enabling global connectivity and fostering political engagement. However, this increased connectivity also introduces risks, including the spread of misinformation, polarization, and the suppression of diverse viewpoints, issues that are central to theories in political science like the Spiral of Silence.

Our research employs complementary methodologies—data analysis of social media and online social experiments—to investigate how digital media facilitates or hinders political expression. First, having built a dataset of approximately 32.5 million comments and replies from videos posted by six prominent US news outlets on YouTube, we studied toxic and insulting behavior in user interactions and its impact on conversation flow and disengagement. We find that toxic and insulting comments are widespread in online political discussions, especially during politically charged periods. Toxic top-level comments often trigger similarly negative replies, creating a self-reinforcing cycle of escalating toxicity. We identify a latent state tied to disengagement, where users become less active but more likely to post toxic content, enabling antisocial behaviors to thrive.

To complement such findings, we designed two computerized, controlled social experiments online. In the first one, participants collaboratively evaluated political content under conditions of overt or covert political affiliations. Results reveal that collaboration improves the quality of fact-checking, but overt political affiliations reduce the quality of outputs, underscoring the benefits of anonymity in collaborative settings. Our work suggests that the anticipation of conflict may lead individuals to conceal their political views, in line with the Spiral of Silence theory.

The second experiment explicitly examines the Spiral of Silence theory, which posits that fear of social isolation and perceptions of majority opinion influence individuals' willingness to express opinions. While we did not find a significant relationship between fear of isolation and the willingness to share opinions, other results support key aspects of the Spiral of Silence theory. Individuals with high communication apprehension were less likely to participate in controversial discussions, while those with high attitude certainty and those who considered an issue highly important were more likely to express their opinions. Our results indicate that individuals rely heavily on their perceptions when deciding whether to express their political opinions or not.

This thesis demonstrates that self-censorship in non-agreeable environments plays a pivotal role in shaping political discourse by discouraging participation. Through empirical evidence supporting the Spiral of Silence theory, and by combining theoretical, experimental, and observational methods, we uncover the complexities of online political conversations. Our work emphasizes the need to account for self-censorship dynamics in social sciences research, offering insights into the interplay between individual behavior and collective dynamics, with an aim to foster healthier, more inclusive digital discourse in modern society.

Acknowledgements

I would like to thank my PhD supervisor, Prof. Gerardo Iñiguez for the support, patience and guidance needed to complete this challenge. I would also like to thank my second supervisor, Prof. Tiago Peixoto, for the disposition to take me on and helping me through the last year of my PhD. I acknowledge and thank my collaborators Prof. Ángel Sánchez, Prof. Júlia Koltai and Prof. Taha Yasseri for guiding me through the work and for being generous with their knowledge, time and patience.

I thank my parents, José and Narcisa, and my siblings, Javier and Julieta, for being a loving, warm, reliable and consistent presence throughout my life. You taught me everything I know about unconditional love, forgiveness and acceptance.

To my colleagues at the Department of Network and Data Science: I admire your talent, integrity, and resilience and solidarity in the face of individual and collective challenges. Above all, I thank you for your friendship. You provided the softness, fun and light-heartedness that I did not know I needed to make it through these years.

A special recognition to my husband, Felipe. I am in awe at your curiosity, talent, tenacity, and above all, humility. Every day, you teach me invaluable lessons on approaching life with an open heart, kindness, and generosity. Having met you and being loved by you has been the greatest joy and the honor of a lifetime.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As digital media become more widely used to engage with political content, online social platforms increasingly cement their roles as key spaces for social interaction and the exchange of ideas [1,2]. This transformation is rooted in the rapid cultural and technological shifts driven by communication technologies, which have reorganized human social networks and integrated them with technological systems [3,4]. These changes have unfolded on an evolutionarily short timescale, outpacing natural selection's ability to adapt our innate behaviors [5]. Bak-Coleman et al. [4] identify four critical areas in which these rapid shifts are impacting social dynamics and reshaping the social landscape.

First, the scale of human social networks has expanded dramatically. Technological advancements, particularly social media, have facilitated global connections involving billions of individuals. While this interconnectedness presents opportunities, it also creates challenges such as heightened competition, difficulties in collective decision-making, and maintaining cooperation [6–8]. The size itself of modern networks introduces complexities that strain social systems.

Second, by enabling frequent interactions across long distances, communication technologies have introduced fundamental changes to network structures. These structural shifts significantly affect the dissemination of information, including the spread of misinformation [9–11]. This connectivity allows geographically distant individuals to interact more frequently, but also heightens the risks associated with network-induced dysfunctions.

Third, information fidelity has improved due to advancements in technology, allowing information to propagate with minimal degradation. While this enhances accessibility, it also facilitates the rapid spread of misinformation [12]. Rapid information flows can overwhelm cognitive processes, resulting in less accurate decisions [13], and repeated exposure to falsehoods within certain societal segments can erode their ability to discern fact from fiction [14].

Finally, algorithmic feedback is reshaping social dynamics by influencing both individual and collective outcomes. Algorithms, designed primarily to maximize attention, content consumption, and thus company profitability, often reinforce biases [15]. We are becoming increasingly reliant on algorithms for information-foraging processes, but these systems lack sufficient incentives to promote informed, just, healthy, and sustainable societal outcomes [16].

By enabling individuals to create, share, and distribute content with unprecedented ease, these platforms have fundamentally reshaped public communication, fostering greater connectivity and interaction. This enhanced connectivity fosters opportunities for political deliberation and engagement, encouraging the exchange of diverse perspectives [17, 18]. However, heightened connectivity is a double-edged sword [19]. Social media platforms have profoundly influenced political discourse, earning these platforms the label of “liberation technology” for their role in empowering civil society movements. Examples include the Arab Spring [20], the Green Movement [21], the “Me Too” movement [22], and Black Lives Matter [23, 24]. However, they have also been implicated in harmful events like the January 2021 attack on the US Capitol [25, 26].

While social media enhances connectivity and provides a platform for diverse viewpoints, enabling political deliberation and fostering depolarization [17, 27], it also introduces significant risks that threaten the potential for productive deliberation and pose substantial challenges to the democratic process. The shift to increased connectivity transfers power to large digital media platforms, which now play a critical role in managing the digital infrastructure where these interactions occur. This influential position allows them to exert substantial control over both the infrastructure and the flow of information within it, raising key questions about data access and platform accountability [4, 28, 29].

Understanding the dynamics and quality of discussions within the interconnected, algorithm-driven digital landscape is essential in this context. Political communication research has thoroughly examined how changes in networked communication technologies shape social movements, institutional politics, and political engagement [30,31]. Although the global shift to digital media has been associated with declining trust in politics [32] and mainstream media [33], as well as with the rise of populism [34], hate speech [35,36], and increasing polarization [37,38], it has also democratized access to information, enhanced political participation [39,40], and has the potential to improve political knowledge [41, 42]. The existing literature presents conflicting views on the influence of digital media on political expression [19]. The observed negative effects are often linked to the Spiral of Silence theory in political science, which asserts that individuals are more likely to express political opinions online if they believe their views are widely shared [43, 44]. This area of research is crucial, as evidence suggests that political expression on online platforms precedes political participation [45].

The Spiral of Silence theory suggests that fear of social isolation and perceptions of public opinion shape individuals’ willingness to express opinions on controversial topics [46]. It posits that individuals continuously monitor their social environment to assess dominant views, which influences their readiness to voice their own perspectives [47]. By emphasizing the tendency to monitor the social landscape, the theory aligns with other frameworks that conceptualize public opinion as a dynamic process [48]. Similarly, the related concept of preference falsification suggests that people may publicly endorse views they do not privately hold when their personal beliefs conflict with those of the perceived majority [49].

This form of social influence is not a recent phenomenon. For instance, experiments by Latané and Darley [50] showed that individuals are more likely to report emergencies when alone than in groups, suggesting that even considerations of physical well-being can be overshadowed by "social fear"—the fear of appearing foolish if their concerns prove unwarranted [51, 52]. This concept aligns with Nisbett and Kunda's [53] suggestion that awareness of holding a minority opinion can lead individuals to avoid controversial topics, aiming to prevent offending those with differing views—an idea central to the Spiral of Silence theory. Moreover, prior research has proposed that presenting oneself in alignment with social norms is a fundamental skill essential for the development and maintenance of complex social structures [54, 55].

While most of these frameworks were intended to describe a pre-digital context, they remain relevant to contemporary issues of self-censorship and opinion dynamics on social media. Social media, in theory, offers a platform for political behavior free from formal constraints. However, substantial evidence demonstrates that significant social pressures continue to shape online interactions. Studies have found that self-censorship is particularly prevalent on social media [56–58], especially among individuals who are highly sensitive to social criticism [59, 60].

The idea that fear of social backlash leads individuals to withhold their public expressions presents a paradox: greater self-censorship should logically result in more uniformity of expressed opinions, yet the persistence of polarization suggests otherwise [61]. Schulz et al. [62] address this apparent contradiction by arguing that, beyond self-censorship, individuals actively and strategically tailor their online expressions to appeal to specific "imagined audiences" [63]. They found that the direction of a person's online political expression is strongly influenced by the perspectives of the accounts they engage with. That is, public expression, shaped by complex social influences [64], cannot be regarded as a simple register of political attitudes, making it an unreliable substitute for survey data [62, 65, 66].

This work aims to understand how digital media may encourage or hinder diverse forms of political expression in online forums, by investigating the relationship between human interaction and processes of opinion expression. To achieve this, we use complementary research methods: the analysis of observational data from online social media platforms (Chapter 2) and data from computerized, controlled social experiments online (Chapters 3 and 4). These approaches offer distinct advantages and effectively address the limitations of the other.

Social media platforms like YouTube provide valuable observational data that can reveal large-scale patterns and trends in collective behavior. However, the influence of platform algorithms, user activity, and content moderation practices introduces challenges related to the generative processes and completeness of the data. These factors make it difficult to infer causal relationships and validate findings [4, 62].

Computerized online social experiments offer a targeted and controlled method for hypothesis testing that complements the observational insights from social media data. These experiments allow researchers to study individual and group behaviors in digital environments under specific conditions. Unlike traditional approaches that focus on either micro-level individual actions or macro-level patterns, online experiments bridge these scales by capturing the emergent dynamics of interactions in real-time [4, 67, 68].

In Chapter 2, we use YouTube comments from the 2020 US presidential elections to examine the prevalence of toxic and insulting language in online political discussions and investigate a potential link between negative sentiment and conversation disengagement. Although online communities in principle allow for constructive discussions in the face of political disagreement, the common occurrence of antisocial behavior—including insults, aggression, and ideological hostility—undermines their role in supporting tolerant deliberation in democratic societies [69, 70]. Harassment is of particular concern. A recent U.S. study shows that although the overall percentage of people experiencing online harassment remained stable at 41% between 2017 and 2021, the incidence of severe harassment increased sharply from 15% in 2014 to 25% by 2021 [71]. According to a related U.S. study, political views are the leading cause of online harassment, with one in five adults reporting harassment based on their political beliefs [72]. Toxic comments, including insults, are the most common forms of harassment encountered online [73]. In the realm of online political discourse, toxicity not only amplifies extreme viewpoints and discourages moderate and marginalized voices, but it also deepens political divides, heightens polarization, and raises safety concerns that deter diverse participation in political discussions [74–77].

Although YouTube is one of the largest and most engaging social media platforms, it has received relatively limited academic attention compared to other social media [78]. Previous research suggests YouTube's role in fostering political radicalization and ideological echo chambers through personalized recommendations, finding that while content generally aligns with users' preferences, a notable right-wing bias leads far-right and moderate users to encounter more ideologically aligned and problematic content [78, 79]. Other work has used YouTube data to identify consistent patterns of toxicity across different platforms and periods, finding that toxicity and user participation in debates operate independently, while diverse opinions among users may contribute to increasing toxicity levels [80]. Building on these efforts, our study provides an in-depth analysis of political conversations on YouTube, expanding the scope to include both toxic and insulting behaviors. We specifically focus on the work of Avalle et al. [80], as their results appear to contradict our working hypothesis. Our study, however, differs from this work in one important aspect: the definition of conversation, which may make our results difficult to compare.

Even though toxicity is a widespread issue in online discussions across various platforms, its reported prevalence varies due to differing definitions and measurement approaches [70,76,80]. We use Google's Perspective API, a machine learning classifier trained on labeled data from sources like Wikipedia and the New York Times, to assess toxicity and insults [81, 82]. Our analysis focuses on two of the seven attributes of abusive comments that Perspective can identify: toxicity and insult. Perspective defines toxicity as a "rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion", and an insult as an "inflammatory, or negative comment towards a person or a group of people". The classifier assigns a probability score between 0 and 1 to indicate how likely a comment contains a given attribute [83].

While widely used, machine learning-based toxicity detection systems like the Perspective API have limitations due to their reliance on accurate and unbiased training datasets, which are challenging and resource-intensive to create [84]. Consequently, it is not surprising that the Perspective API may struggle to accurately identify and address toxicity in all contexts. For example, it has been found to assign higher toxicity scores to comments containing terms associated with frequently targeted groups (e.g., "Black", "Muslim", "feminist", "woman", "gay") [85]. Additionally, the classifier's performance can vary by language; one study found that German content received significantly higher toxicity ratings, leading to nearly four times more moderated comments than in their English translations [86].

Despite these concerns, automated toxicity detection systems remain the most practical approach for large-scale analyses of social media data [80, 84, 87]. Research has shown that the Perspective API performs comparably to human coders in classifying toxic comments on platforms such as Reddit [88], and Facebook and Twitter [76, 89]. Our examination of examples of comments at various levels of the toxicity and insult scales suggests that, while not perfect, the Perspective API's toxicity and insult scores generally correspond well with the content of the comments.

We built a dataset of approximately 32.5 million top-level comments and replies from videos posted by six prominent U.S. news outlets on YouTube, chosen for their active comment sections and diverse political ideologies. Our dataset includes content from three left-leaning sources—ABC News, CBS, and CNN—and three right-leaning ones: Newsmax, Fox News, and One America News Network. To focus on politically relevant discussions, we chose a study window-September 2020 to April 2021-to cover key events surrounding the 2020 U.S. presidential election and the first 100 days of Biden's administration. We observed significant spikes in toxicity and insults coinciding with major political moments such as the Black Lives Matter protests, Election Day, and the Capitol riots of January 6. Furthermore, toxic top-level comments are more likely to elicit similarly negative replies, suggesting a cyclical relationship where toxicity fuels further toxicity. Using Hidden Markov models, we study individuals' responses when confronted with a toxic and insulting environment. Our findings suggest that, in some contexts, toxic and insulting posts tend to surface toward the end of discussions. These findings provide insights into the complexities of online political deliberation and emphasize the importance of considering self-censorship dynamics in understanding digital discourse.

Although informative, this study has important limitations in regards to the completeness and quality of the data. We assume that all conversations end due to toxic or insulting behavior. However, people leave conversations for reasons other than toxicity. Therefore, we might be over-estimating the effect of negative sentiment on disengagement. To address this issue, we designed two online experiments to more accurately measure disengagement. In Chapter 3, we describe the design of an online experiment to test an approach that implements user collaboration to label and annotate online political content.

Misinformation and disinformation permeate all media forms but are notably exacerbated on social media platforms due to their rapid content dissemination [90]. Allcott et al.'s study of fake news during the 2016 U.S. election revealed that the average American encountered between one to three posts from known fake news publishers [91]. False information spreads faster and wider than truthful content, with automated users like bots amplifying posts related to elections [92]. Furthermore, interaction between individual exposure and confirmation bias exacerbates the impact of misinformation, as people are more likely to believe content aligning with their political views [93, 94]. Misleading content—defined as factually accurate but lacking context such as satire, citizen journalism, and one-sided reporting—presents another challenge for content moderators. Its prevalence on social media platforms is attributed to character limitations and the broadcast-like nature of platforms like Twitter [95–97]. Identifying misleading content is considerably more challenging compared to misinformation and disinformation, posing a threat to the quality of political deliberation online.

To tackle these challenges, social media platforms have experimented with various strategies, ranging from professional fact-checking to leveraging the collective knowledge of users for content moderation [98]. While traditional fact-checking, relying on human experts for content review and truthfulness classification, can be effective, it's neither scalable nor cost-efficient [99]. Thus, alternatives such as automated methods, crowd-sourcing, or hybrid approaches show promise in addressing this issue [100]. Recently, platforms have begun experimenting with community-based fact-checking solutions like X's Community Notes. These approaches capitalize on the proven effectiveness of collaborative problem-solving, as opposed to crowd-sourced approaches where participants work individually and results are later combined [101–103].

We conducted a online experiment to assess the efficacy of community-based approaches in combating misinformation and disinformation. In the online experiment, participants were assigned the task of composing individual and collaborative notes for 40 political tweet, sourced from Democrat and Republican accounts. Participants, recruited via Prolific, were randomly paired with partners from either the same or different political affiliation, resulting in three group configurations: (1) Democrat-Democrat (DD), (2) Republican- Republican (RR), and (3) Democrat-Republican (DR). Moreover, each team was randomly assigned to one of two treatments: in Treatment 1 (the 'Overt' treatment), participants could view each other's political affiliations, while in Treatment 2 (the 'Covert' treatment), participants were unaware of each other's affiliations.

Our results indicate that teams, on average, produce more helpful notes than individuals, supporting our first hypothesis about the positive impact of collaboration on fact-checking. Interestingly, the DD (Democrat-Democrat) team outperformed both RR (Republican-Republican) and DR (Democrat-Republican) teams in providing helpful notes on Democratic tweets. This finding contrasts with our expectation that heterogeneous groups (DR) would outperform homogeneous groups (DD and RR). Additionally, the quality of collaborative notes declined when participants were explicitly aware of each other's political affiliation. In contrast, when political affiliations were covert, teams produced higher-quality notes, demonstrating the benefits of anonymity in collaborative fact-checking.

Although the work in Chapter 3 does not directly examine the dynamics of political expression in online forums, which is the central focus of this dissertation, it contributes to the study in two important ways. First, the experiment establishes a context in which political expression is a prerequisite for completing a specific task. By doing so, it compels participants to consider their political views when interacting with their partners. From the perspective of opinion diversity, heterogeneous groups were initially expected to outperform homogeneous groups. Our results challenge this assumption: while teams with overt political affiliations did not perform as well, those with covert affiliations produced better outcomes. This suggests that the anticipation of conflict may lead individuals to disengage from the task or conceal their political views, consistent with the Spiral of Silence theory. Second, the study served as a testing ground to refine the design of interactive experiments, particularly in terms of coding implementation and participant recruitment. The lessons learned from this work provided valuable insights that informed the experiment in the following chapter.

In Chapter 4 we designed an online experiment to explicitly test the Spiral of Silence theory. In particular, we investigate whether online settings that incentivize connections with others heighten individuals' fear of isolation, thereby making them less likely to share their true opinions on controversial topics, particularly when their views differ from those of others. Specifically, we seek to experimentally test the main premise behind the Spiral of Silence theory [46].

Before the advent of social media, people's interactions were constrained by physical and geographical boundaries. Social media has revolutionized communication by shifting at least some of these interactions online. In doing so, it has facilitated connections among like-minded individuals, decentralized the distribution of information [104], and created a digital record of these interactions [105]. Understanding how individuals embedded within interaction and communication networks exercise and respond to social influence is crucial for studying collective civic behavior. To examine how context shapes civic behavior, it is essential to understand what information people encounter in their daily interactions, where they encounter it, and how it informs their beliefs and actions [106].

This research investigates the relationship between human interaction and the processes of opinion formation and expression. Established models of opinion dynamics often assume that interactions between individuals with divergent opinions primarily influence whether or not they change their views (for a comprehensive review, see [107]). While such simplifications are necessary to make the modeling process more manageable, they overlook the fact that sensitivity to others' feedback does not always lead to reconsideration of one's opinion. Instead, it may affect a person's willingness to express their opinion. Positive feedback tends to increase motivation to share opinions publicly, while negative feedback often discourages it [104].

In an era where an increasing number of people engage with political content online, understanding how individuals respond to feedback and how their choices influence others has become more crucial than ever. Theories such as the Spiral of Silence and preference falsification suggest that fear of social backlash drives individuals to withhold or modify their public expressions [46]. The theory further suggest that to avoid being perceived as siding with the unpopular position in a public debate, individuals first evaluate the opinion climate within their immediate social context [108]. Those who perceive themselves as holding a minority opinion are likely to remain silent, while those who believe they are part of the majority are more inclined to voice their views [104]. This process creates a self-reinforcing cycle in which individuals with minority opinions stay quiet, perpetuating a false impression of consensus or conformity [62]. Conversely, a minority opinion can sometimes achieve majority status when its supporters are well-connected within the social network, while holders of the majority opinion are less connected. Under such conditions, those with the majority view may erroneously perceive themselves as being in the minority and, as a result, refrain from expressing their opinions [104].

This work builds on the studies by Gainsbauer, Olbrich, and Banisch [104] and Porten-Cheé and Eilders [109] to investigate whether collaborative online environments can heighten individuals' fear of isolation, thereby decreasing their willingness to publicly express opinions on controversial topics. Gainsbauer, Olbrich, and Banisch [104] proposed a modeling framework combining game-theoretical and dynamical systems approaches to analyze the emergence of a spiral of silence. They concluded that an "internally well-connected minority community" is the only necessary condition for this phenomenon, while mass media influence, though not essential, can still amplify the process by failing to objectively represent diverse viewpoints [104].

There is ample literature on experimental evidence for the micro-mechanisms underlying a potential spiral of silence. For a comprehensive review see Matthes, Knoll and von Sikorsky [44]. Although most of this work has concentrated on studying the implications of face-to-face interactions, there are some examples in the literature of work that aims to study the effect of user-generated content in the internet in general and social media in particular. Porten-Cheé and Eilders [109] examined how social media affects individuals' willingness to speak publicly, focusing on the roles of user-generated content, mass media, and minority opinion status. Their findings did not strongly support the Spiral of Silence theory in online settings, although they acknowledged potential biases in their sample and the relatively uncontroversial status of climate change in Germany, where the study was conducted.

This thesis examines how the fundamental human need for connection influences individuals' willingness to share opinions on controversial topics in online settings. Through an online experiment, our research explores whether environments that encourage collaboration increase fear of isolation, potentially discouraging participants from expressing their genuine opinions on contentious issues. Our central research question is: Can collaborative online environments heighten fear of isolation, thereby reducing individuals' willingness to publicly share controversial opinions?

The experiment encourages participants to form and maintain as many connections as possible, hypothesizing that incentives will prompt participants to carefully manage which opinions they disclose. By providing experimental evidence for the Spiral of Silence hypothesis in an online setting, our research addresses a gap in the literature. Furthermore, this thesis seeks to bridge theories of political communication with insights from network science, offering a novel contribution to understanding how online community dynamics impact political deliberation.

When studying the Spiral of Silence theory, the key dependent variable is the willingness to express an opinion, which is directly influenced by the fear of social isolation and perceptions of majority climate [44, 110, 111]. To navigate this fear, individuals monitor their social environment using a "quasi-statistical sense" to gauge prevailing opinions [108].

The effect of fear of isolation on the willingness to express opinions is moderated by various factors, three of which are explicitly tested in this experiment: (1) communication apprehension [110], (2) attitude certainty [111], and (3) issue importance [109]:

- **Communication apprehension** refers to an individual's discomfort or anxiety associated with real or anticipated oral communication [110]. In simpler terms, it reflects an individual's level of shyness. It is expected that individuals with high communication apprehension will be less likely to engage in discussions on controversial topics and, consequently, less likely to share their opinions.
- **Attitude certainty** is defined as the extent to which an attitude is resistant to change

and predictive of behavior [111]. Individuals with high attitude certainty are less fearful of social isolation and more inclined to share their opinions publicly. Notably, research has shown that a spiral of silence primarily affects those with low to moderate levels of attitude certainty [111].

- **Issue importance** is related to, but distinct from, attitude certainty. While attitude certainty pertains to how confident an individual is in their position, issue importance refers to the significance one assigns to a specific topic [109]. Individuals who consider a topic highly important are generally more likely to express their opinions on the issue. However, high issue importance does not always equate to high attitude certainty. Regarding the willingness to speak, individuals who perceive an issue as important are more likely to share their opinions with others.

We find that while fear of isolation does not significantly influence the odds of sharing, the interaction between fear of isolation and willingness to share is significant only when an issue is deemed important to others; for less important issues, fear of isolation encourages greater sharing. Our findings demonstrate that individuals consider the prevailing opinion climate when deciding whether to voice their views or not [110], those with strong attitude certainty are more inclined to share publicly [111], and perceiving an issue as important increases the likelihood of sharing opinions [109]. These results align with key aspects of the Spiral of Silence theory.

1.0.1 List of publications

1. Juncosa, G., Yasseri, T., Koltai, J., and Iñiguez, G. “Toxic behavior silences online political conversations”. *Submitted for publication to EPJ Data Science*. arXiv preprint arXiv:2412.05741. (2024).

2. Juncosa*, G., Mohammadi*, S., Samahita, M., and Yasseri, T. “Teams work better at evaluating and annotating misleading political content”. *Submitted for publication* (2024).

3. Juncosa, G., Sánchez, A., Koltai, J., Yasseri, T., and Iñiguez, G. “Exploring the Impact of Online Collaborative Environments on the Willingness to Express Opinions: A Spiral of Silence Perspective”. *In preparation* (2024).

Chapter 2

Toxicity in online political deliberation: YouTube during the 2020 US presidential elections

2.1 Introduction

Understanding how individuals respond to social influence is essential for studying collective political behavior in online spaces. While many studies on public forum opinions focus on social feedback [106], they often overlook how human interactions can lead to self-censorship [62, 104]. This chapter explores political deliberation in online environments, investigating the hypothesis that individuals may withhold minority opinions in public due to fears of encountering toxic behavior.

In this chapter, we use YouTube and the 2020 U.S. presidential election as a case study to examine the prevalence of toxic and insulting content in online political discourse. Building on prior research into YouTube conversations [78–80], we investigate how such behaviors shape the dynamics of political discussions, with a particular focus on toxicity and insults, and their impact on conversation dynamics.

To explore these patterns, we compile a dataset of approximately 32.5 million comments and replies from videos posted by six prominent US news outlets on YouTube. We have selected these outlets for their active comment sections and diverse political perspectives, comprising three left-leaning sources (ABC News, CBS, and CNN) and three right-leaning sources (Newsmax, Fox News, and One America News Network). To capture politically relevant discussions, we further focus on the 2020 US presidential election. The analysis period begins in September 2020—just over 60 days before Election Day—and extends through April 2021, covering president Biden's first 100 days in office. This period captures politically charged discussions leading up to the election and during the early days of the new administration, a time traditionally marked by heightened media attention.

We analyze both individual comments and the conversations they form, defining a conversation as a time-bound exchange of replies to a single comment. Replies posted within 10 days of the original comment are included, assuming users directly respond to it, even if they mention other people in the reply text. This approach yields over 2.9 million conversations.

We find that toxic and insulting comments are prevalent in online political discussions, particularly during periods of heightened political tension. Furthermore, toxic and insulting top-level comments are more likely to provoke similarly negative replies, indicating a cyclical relationship where toxicity perpetuates further toxicity, and insults provoke more insults. Finally, through hidden Markov models, we identify a latent state linked to toxicity-driven disengagement. This state is characterized by reduced user activity and an increased likelihood of posting toxic content, fostering an environment where extreme and antisocial behaviors prevail. This finding is confirmed for insulting behavior and across different video ensembles.

2.2 Methods

2.2.1 Hidden Markov Models

A Hidden Markov Model (HMM) is a statistical framework designed to analyze systems where the observed data is generated by underlying states that are not directly visible. As a specific form of a dynamic Bayesian network [112], HMMs excel at modeling temporal and sequential data. First introduced in speech recognition, HMMs have since been successfully applied to various fields, including biological sequence analysis, handwriting recognition, and protein-protein interaction predictions [112]. These models operate on the principle of the Markov process, where the future state depends solely on the present state, not the sequence of past states. However, unlike traditional Markov models, HMMs introduce hidden states, meaning the actual states of the system cannot be observed directly. Instead, only observable symbols or events generated by these states are accessible [113, 114].

An HMM consists of several essential components. Hidden states form the backbone of the model, representing the unobservable conditions influencing the observed data. Observable data, such as a sequence of symbols or events, serves as the manifestation of these hidden states. The model incorporates transition probabilities, which define the likelihood of moving from one hidden state to another, and emission probabilities, which specify the likelihood of observing particular data given a specific hidden state. Together, these parameters enable the HMM to model the relationship between observable data and the hidden processes generating it [112–114].

Training an HMM involves estimating these probabilities to maximize the likelihood of observed sequences, typically using iterative algorithms such as Expectation-Maximization. Once trained, the model is capable of performing several critical tasks. It can evaluate sequences to determine how well they align with the training data, decode observed sequences to identify the most likely sequence of hidden states, and even generate new sequences that mirror the patterns of the training data. Despite its efficiency in solving these problems, finding the ”best” set of probabilities is computationally challenging, as it involves solving an NP-hard optimization problem [112].

HMMs are particularly useful for applications requiring probabilistic modeling of unobservable dynamics. For instance, in speech recognition, HMMs map sound patterns to spoken words, while in biological research, they predict gene sequences or protein interactions based on observed biological data [112–114]. A conceptual example of their use involves inferring weather conditions from observed activities, such as beach visits or staying indoors. In this case, while the actual weather remains hidden, the HMM can deduce the most probable weather patterns based on the observed behaviors [115].

This work leverages the utility of Hidden Markov Models (HMMs) as a robust analytical framework for making inferences about systems with hidden dynamics. We apply HMMs to examine the relationship between negative sentiment and disengagement in conversations under YouTube videos. Specifically, we them to *learn* model parameters in various contexts. We infer the transition and emission probabilities that best align with the sequences of comments, or conversations, in our dataset, tailoring the parameters for each of the six channels. Our approach focuses on learning these parameters for groups of videos, allowing us to capture and analyze aggregated behavior across video collections. We organize videos in two distinct ways: (1) by grouping them according to their news media channel publishers, and (2) by categorizing them based on the topics they address, as identified through topic modeling, regardless of their channel of origin. We use these two alternative methods for defining video ensembles to ensure that the observed relationship between negative sentiment and disengagement is not dependent on the specific group definitions.

To achieve this, we fitted a two-level HMM, as illustrated in Figure 2.1. In this model, $X_1 = 0$ represents the absence of activity or the conclusion of a conversation, $X_2 = 1$ corresponds to a non-toxic or non-insulting post, and $X_3 = 2$ signifies a comment categorized as either toxic or insulting. To convert our conversation threads into a sequence of observations, we began by classifying each comment as either toxic or insulting ($X_3 = 2$) or non-toxic/non-insulting ($X_2 = 1$). This process transforms a chronologically ordered text sequence into a sequence of 1s and 2s. Additionally, we appended a zero ($X_1 = 0$) to the end of every sequence/conversation to indicate its conclusion, ensuring that all conversations end with a 0.

Due to the size of our dataset, a sampling process was necessary, since fitting the model param-

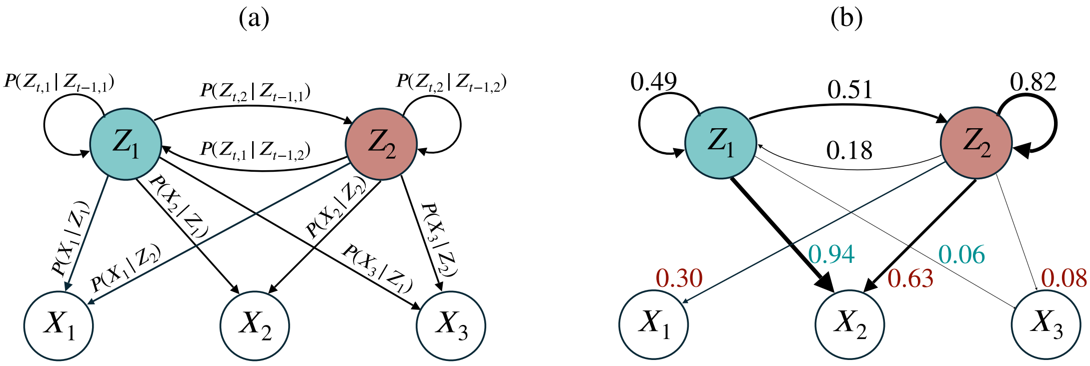


Figure 2.1: **Diagram of a 2-state hidden Markov model and inferred transition and emission probabilities.** (a) Hidden Markov model describing the relationship between unobserved states and observed data over time. The model has two latent states (Z_1 and Z_2) and three observations (X_1, X_2 and X_3). The arrows between Z_1 and Z_2 represent the transition probabilities between the two latent states ($P(Z_{t,i}|Z_{t-1,i})$ where $i = 1,2$). The top-to-bottom arrows indicate the emission probabilities of observations given the latent states ($P(X_j|Z_i)$ where $j = 1,2,3$). Panel (b) shows the inferred transition and emission probabilities for a 2-state Hidden Markov Model including conversations of all lengths. The shown probabilities are averages over multiple model fits, with standard errors consistently below 0.001. Additional checks confirm that filtering short conversations has no significant impact on inferred probabilities (see Fig. A.4).

eters for most ensembles required too many conversations to be processed at once. Through experimentation, we determined that no more than 45,000 conversations could be used at a time for training. On average, fitting the model parameters for one realization required approximately two hours. To address this, we sampled 45,000 conversations with replacement for each realization, splitting each sample into a training set (80%, or 36,000 conversations) and a test set (20%, or 9,000 conversations). We conducted as many realizations as our computational resources allowed, and Table A.3 in Appendix A.4 provides a summary of the number of model fits included in our results.

2.3 Results and discussion

2.3.1 Toxic and insulting behavior appears across online political conversations

We use YouTube as a case study to explore the prevalence of toxic and insulting content in online political discussions. To focus on politically relevant discussions, we chose to study the 2020 U.S. presidential election. Our analysis period begins in September 2020—just over 60 days before Election Day—and extends through April 2021, covering Biden's first 100 days in office. This time frame captures a broad range of politically charged discussions both leading up to the election and during the early days of the new administration, a period traditionally marked by heightened media attention.

We selected the channels for our dataset by first analyzing media outlets' content and their engagement metrics, as listed in the 2019 AllSides Media Bias Chart [116]. The AllSides Media Bias Meter™ rates media outlets on a scale from -6 to +6, where 0 represents the Center, -6 indicates the farthest Left, and +6 denotes the farthest Right. This rating system replaces the previous five-category classification (Left, Lean Left, Center, Lean Right, and Right) to capture a more detailed view of media bias. Bias scores are determined using a combination of methods, including editorial reviews by a multipartisan panel and blind bias surveys, where participants rate outlets based on sampled content [117]. We use the bias score to sort channels from left to right, an ordering that will be applied across all figures where applicable.

We narrowed our focus to six outlets with active comment sections that represent a broad spectrum of political ideologies: three left-leaning sources—ABC News, CBS, and CNN—and three right-leaning channels—Newsmax, Fox News, and One America News Network (OAN) (see Appendix A.2 for information on bias ratings). Using the YouTube API, we collected top-level comments and replies on videos posted by these channels during our analysis period, along with engagement statistics for each video. The resulting dataset includes 18,627 videos and over 32 million comments and replies (see Table 2.1).

We examine both individual comments and the conversations they form. A conversation is defined as a coherent, time-bound exchange of posts centered around a single top-level comment and its replies, capturing user interactions centered on a specific topic. Each conversation begins with a top-level comment posted on a YouTube video, which serves as the prompt for a conversation. All replies to this top-level comment are included in the conversation, ordered chronologically. To maintain relevance and focus, only replies posted within 10 days of the top-level comment are considered part of the conversation.

YouTube supports only first-level replies, but users often work around this limitation by tagging or mentioning others in their comments to continue the conversation. Using these mentions, it is therefore possible to reconstruct sub-threads; however, this is beyond the scope of this work. For the sake of simplicity, we assume that every reply targets the top-level comment, even if it mentions someone other than the comment's author. Consequently, each video contains multiple conversations, resulting in over 2.9 million conversations in our dataset. For detailed descriptive statistics, see Table 2.1.

To assess toxicity and insulting behavior, we use Google's Perspective API, a machine learning classifier trained on labeled data from sources like Wikipedia and the New York Times [81,82]. The Perspective API uses machine learning to detect abusive comments by scoring phrases based solely on their textual content, without considering emojis or images [83]. The model evaluates various attributes, including toxicity, insults, severe toxicity, identity attacks, threats, profanity, and sexually explicit content. However, we focus on toxicity and insults, as these attributes are the most prevalent in our dataset (see Figure A.2 in Appendix A.3 for the prevalence of all attributes). Toxicity is defined as a "rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion", and insult as an "inflammatory, or negative comment towards a person or a group of people" [118].

The classifier assigns a probability score between 0 and 1 to indicate how likely a reader would perceive the comment as containing a given attribute. A higher score reflects a greater likelihood of the attribute being present. For example, a toxicity score close to 0 suggests the comment is unlikely to be toxic, a score of 0.5 indicates uncertainty , and a score of 1 means the comment is highly likely to be toxic [119]. Following previous research, we set our classification threshold at 0.6, and conducted robustness checks with different thresholds [80, 120]. While current automated toxicity detection systems face limitations mainly due to training dataset biases [84–86], they remain an effective tool for conducting large-scale analyses at the level of entire populations [76, 80].

We found that the overall prevalence of toxicity (10.1%) and insulting language (10.4%) in our dataset is relatively low. Yet, toxicity levels are higher than those reported in other YouTube studies, which report a prevalence of 4-7% [80]. The higher prevalence of toxicity in our dataset may be attributed to the focus on politically charged topics, which tend to attract more emotionally charged discussions [121]. Additionally, differences in sampling periods and the commenting culture of selected channels may contribute to these differences.

There is also significant variation across channels (Table 2.1). Left-leaning ABC News stands out with the highest level of toxicity (18.7%) and insults (19.5%), suggesting that its comment section may be among the most contentious. In contrast, CBS and right-leaning OAN show the lowest levels of toxic comments (5.6% and 6.3%, respectively) and insulting language (6.0% and 5.9%, respectively). A general trend emerges where right-leaning channels have lower average percentages of toxic and insulting comments compared to left-leaning channels.

2.3.2 Online activity and toxic behavior surge during politically charged events

Moments of heightened national tension, such as controversial legal decisions, incidents of police violence, and election-related events, appear to trigger surges in negative online sentiment. Figure 2.2 shows time series data on the fraction of toxic and insulting posts (blue lines), daily comment counts (gray line), and a 7-day rolling average (red line). Key events are annotated with black dots. The proportion of toxic and insulting comments is defined as the fraction of comments with sentiment scores exceeding 0.6. Initially, both toxicity and insult trends exhibit relatively high averages, possibly reflecting social unrest and widespread demonstrations against racial injustice, particularly those connected to the Black Lives Matter movement. No-

Table 2.1: **YouTube Dataset Descriptive Statistics.** The dataset contains top-level comments and replies from videos posted by six prominent U.S. news outlets chosen for their active YouTube comment sections and for representing a broad spectrum of political ideologies. The time frame begins in September 2020, and extends through April 2021.

	Channel	Bias Score[1]	Videos	Conversations	Comments	Toxicity[2]	Insult[3]
Lean Left	ABC News	-2.40	4314	453290	4020626	18.7	19.5
	CBS	-1.50	4950	361727	2957079	5.6	6.0
	CNN	-1.30	1856	895584	10153723	10.2	9.5
Right	OAN	3.10	1747	49342	613116	6.3	5.9
	Newsmax	3.28	1752	189679	2844506	7.0	8.0
	Fox News	3.88	4008	961907	11890276	9.2	9.9
	ALL		18627	2911529	32479326	10.1	10.4

1. https://www.allsides.com/media-bias/media-bias-chart. Retrieved August 21, 2024

2. Percentage of toxic comments. A comment is toxic if its toxicity score is greater than 0.6.

3. Percentage of insulting comments. A comment is insulting if its insult score is greater than 0.6.

tably, there is a peak in toxicity (Figure 2.2, panel (a)) and a more modest increase in insults (panel (b)) and activity (panel (c)) following the grand jury's decision to charge only one officer in the case of Breonna Taylor's death. This decision was followed by mostly peaceful nationwide demonstrations over the weekend of September 27–28, 2020 [122].

On April 11, 2021, Daunte Wright was fatally shot by police officer Kimberly Potter, an event that sparked protests across the U.S. and globally [123]. These demonstrations overlapped with the ongoing investigation into George Floyd's death at the hands of police officer Derek Chauvin. The timing of Wright's death and Chauvin's sentencing correlates with a rise in both toxicity and insults (panels (a) and (b)), alongside a modest increase in activity (panel (c)).

Following Election Day, online activity continued to climb, with two notable peaks (panel (c)). The first spike coincided with Biden's projected victory announcement, while the second occurred as Trump openly questioned the election's legitimacy. This latter moment also saw an increase in toxic and insulting posts (panels (a) and (b)). The dataset's highest peak in activity appeared on January 6, 2021, the day of the Capitol Attack by Trump supporters, which marked the largest spike in toxicity and one of the highest in insulting posts. In February 2021, another distinct peak in insulting comments appeared, likely linked to events surrounding the COVID-19 pandemic, including vaccination roll out and debates over school reopening [124].

Following Election Day, online activity continued to climb, with two notable peaks (panel (c)). The first spike coincided with Biden's projected victory announcement, while the second occurred as Trump openly questioned the election's legitimacy. This latter moment also saw an increase in toxic and insulting posts (panels (a) and (b)). The dataset's highest peak in activity

appeared on January 6, 2021, the day of the Capitol Attack by Trump supporters, which marked the largest spike in toxicity and one of the highest in insulting posts.

In February 2021, another distinct peak in insulting comments appeared, likely linked to events surrounding the COVID-19 pandemic, including vaccination roll-outs and debates over school reopenings [124]. The pattern of rising toxic and insulting language in response to significant socio-political events suggests a reactive trend, with public discourse intensifying both in volume and negativity after high-profile events that elicit strong emotional reactions. These findings underscore how such events impact online discourse, revealing a cycle where contentious incidents not only attract increased public attention but also heighten the intensity and negativity of conversations.

2.3.3 Toxicity and insults fuel ongoing negativity in online conversations

This section examines the relationship between toxicity and insult scores for replies (Y) and the scores of their corresponding original comments (X). Figure 2.3 illustrates the probability density distribution of toxicity and insult scores for replies (dotted lines in Fig. 2.3a–b). Figure 2.3c–d shows the probability of a reply having a specific toxicity or insult score, conditional on the corresponding score of the comment. Both distributions are skewed to the left, peaking at low scores (below 0.2), suggesting that while both toxic and insulting top-level comments are likely to receive similarly negative responses, replies tend to have lower insult scores than toxicity scores. Areas under the curve for scores exceeding 0.6 are similar for both toxicity and insult, aligning with prevalence estimates of 10.1% for toxicity and 10.4% for insult, as shown in Table 2.1.

In Fig. 2.3, we also examine the probability density functions of toxicity and insult scores for replies (Y) conditional to the scores of their comments (X). The solid lines show the conditional probabilities $P(Y|X)$ across different ranges of X. The shapes of the conditional probabilities suggest that while reply sentiment scores correlate with the sentiment scores of the top-level comments they respond to, toxicity exhibits a wider distribution, whereas insults are more concentrated around lower values, staying closer to zero. Furthermore, we found that the toxicity score of a top-level comment had a significant positive effect on the toxicity score of its replies ($\beta = 0.1366$, $CI = [0.1357, 0.1375]$, $p < 0.001$). Similarly, the insult score of a top-level comment showed a significant positive effect on the insult score of its replies ($\beta = 0.1511$, $CI = [0.1502, 0.1520]$, $p < 0.001$).

In Fig. 2.3a, the blue line represents top-level comments with low toxicity scores (0 to 0.5), indicating that replies to less toxic comments are more likely to have low toxicity scores as well. This pattern is further illustrated in Fig. 2.3c, where $P(0 < Y < 0.5|0 < X < 0.5)$ is higher than $P(0 < Y < 0.5|X)$ when conditioning on greater values of X. In contrast, the yellow and

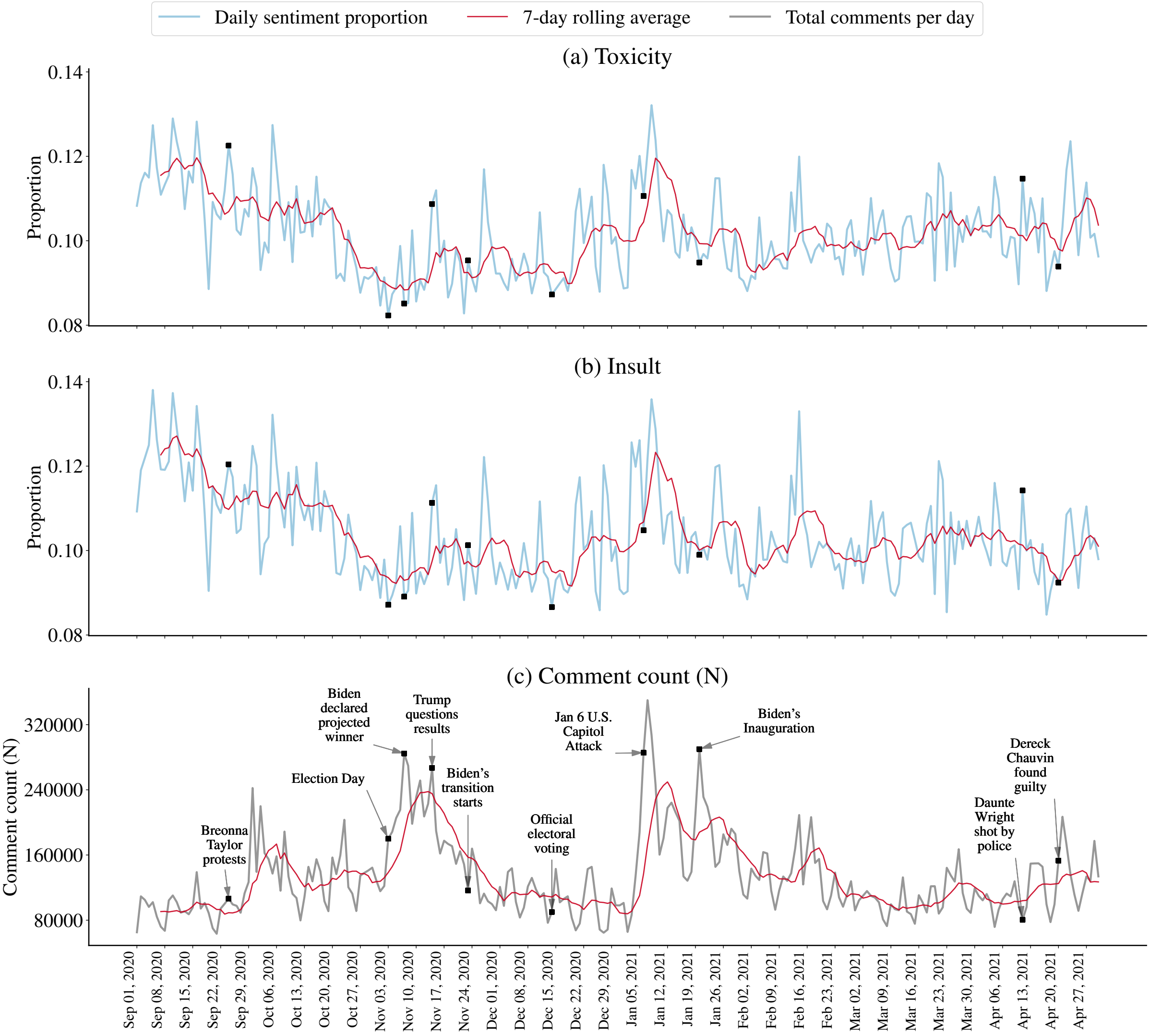


Figure 2.2: **Trends in comment activity and toxicity over time.** Daily sentiment proportions (blue line), total number of comments per day (gray line), 7-day rolling average (red line), and relevant events annotated with black dots. **(a)** Proportion of toxic comments, defined as those with a toxicity score greater than 0.6. Peaks in toxicity are observed around September 2020, coinciding with heightened social and political tensions following the Black Lives Matter protests. A subsequent peak occurs in January 2021, likely linked to the increased tensions surrounding the Capitol attack. (b) Proportion of insulting comments, defined as those with an insult score greater than 0.6. The trend in insulting comments resembles that of toxicity; however, there is a more pronounced peak in February 2021, likely related to events surrounding the COVID-19 pandemic. (c) Total number of comments per day. A notable surge in activity coincides with the controversial events of the 2020 election; yet this increase is not mirrored in the trends for toxicity or insulting comments. The highest comment counts during the January 2021 Capitol attack coincide with spikes in both toxicity and insults.

light blue lines represent top-level comments with intermediate toxicity levels, indicating that as the toxicity of the initial comment rises, so does the probability of a toxic reply. This trend is further highlighted by the brown line, which shows that replies to highly toxic comments are more likely to reach high toxicity levels themselves, surpassing other conditional probabilities when $Y > 0.65$. This pattern holds for insulting scores as well, as shown in Fig. 2.3b. These findings hint at an influence of initial comment sentiment on the nature of replies, with elevated toxicity or insult scores in top-level comments associated with higher likelihoods of similarly negative replies.

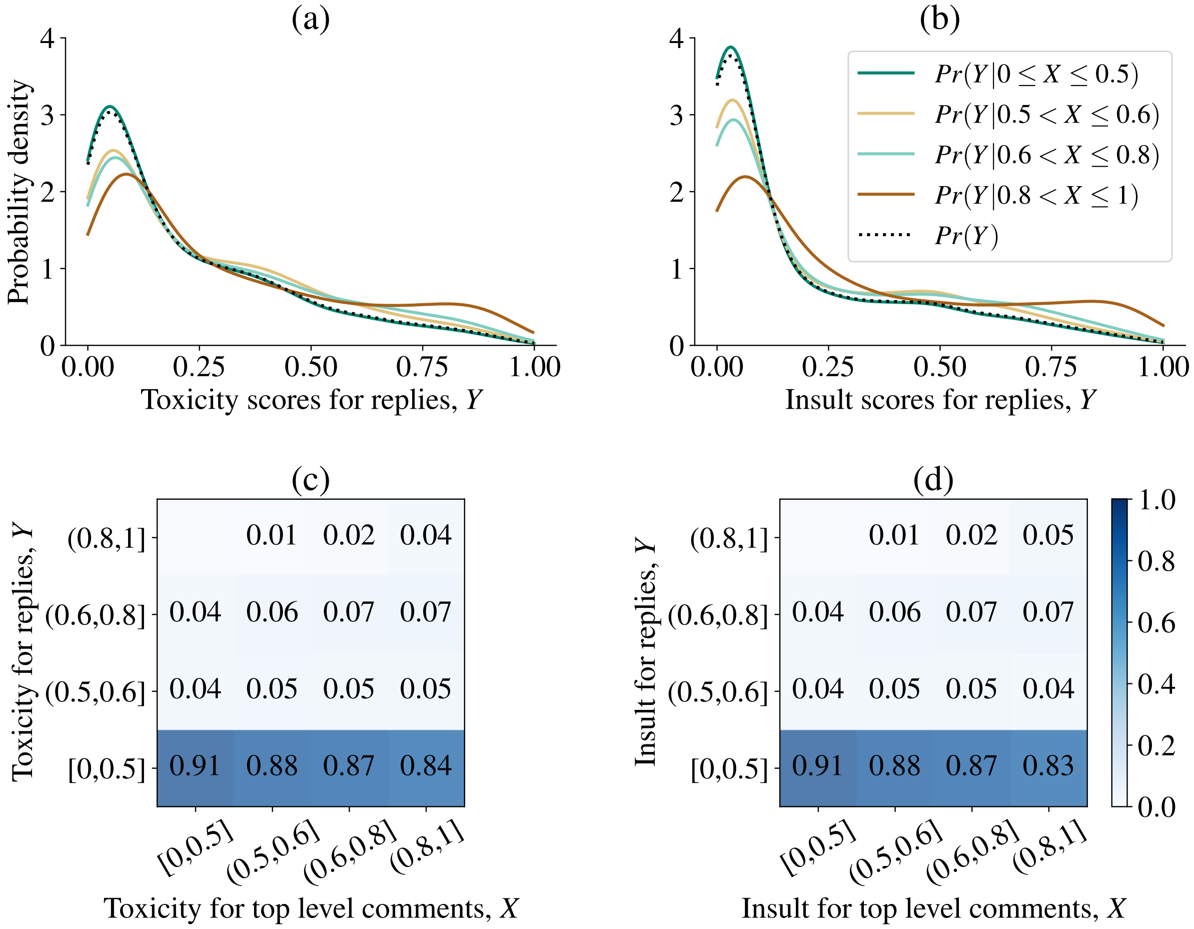


Figure 2.3: **Prevalence of toxicity and insults in YouTube comments and their replies.** (a–b) Conditional probability density functions $P(Y|X)$ of toxicity and insult scores of replies (Y) based on the corresponding scores of their original comments (X). The dotted line corresponds to the probability distribution $P(Y)$, the overall score distribution for replies. (c–d) Estimated probabilities of a reply having a specific sentiment score, conditioned on the sentiment score of the top-level comment. Results for both toxicity and insults show a correlation between the negative sentiment of comments and their response, with replies to highly toxic/insulting comments (brown lines) being more likely to have high toxicity/insult scores (greater than 0.8) than other replies.

2.3.4 Uncovering the role of toxicity via latent variable modeling

While individual comments offer valuable insights, analyzing whole conversations is essential for understanding user interactions online. Our goal is to examine how conversations evolve, particularly whether toxicity can be linked to disengagement. One way to approach this is to consider toxicity as a reflection of an underlying, unobserved conversational state. For instance, a conversation may either foster or hinder dialogue (two possible states), with each state potentially involving toxic or non-toxic content, or leading to the conversation's end. While these states are not directly observable, one can theorize a relationship between expressions of toxicity and a conversation's conduciveness to dialogue.

Here, we use hidden Markov models (HMMs) [113–115] to infer conversational states based on observed patterns of toxic behavior online. Since insult and toxicity are closely related in practice, we further focus exclusively on toxicity. By identifying hidden states, HMMs can provide insights into how toxic and non-toxic posts influence conversational dynamics. From this perspective, conversations progress as sequences of toxic and non-toxic exchanges over time. Their sequential nature suggests transitions between different latent states, which, though unobservable, manifest in distinct sentiment patterns. We allow toxic content to appear in both states (conducive and non-conducive to dialogue) rather than restricting it to one, enabling us to determine where toxicity is more likely to occur.

The utility of Hidden Markov Models (HMMs) is best understood through examples. Consider a future climatologist studying historical global warming trends to reconstruct past weather patterns without direct records. Instead of weather data, the climatologist uncovers a diary detailing daily ice cream consumption. Recognizing the relationship between weather and ice cream consumption, the climatologist can employ HMMs to analyze this observable behavior and infer the underlying weather conditions that likely influenced it [115]. Similarly, in our study, we may lack direct information about the specific states a conversation undergoes. However, by observing external indicators—such as whether a comment is toxic or insulting—we can infer the latent states shaping the conversation.

Formally, a Hidden Markov Model consists of a sequence of observed variables, denoted as X, and a corresponding sequence of hidden states, denoted as Z. The observed variables X represent data we can directly measure, such as the sentiment of a comment. The hidden states Z, however, are latent variables that represent the underlying conditions (e.g., conversational tone) influencing the observations. Each hidden state Z generates an observation X according to a probability distribution, and transitions between hidden states follow a Markov process, where the probability of transitioning to a new state depends solely on the current state. This framework allows us to model the progression of sentiment in conversations, capturing how hidden conversational states influence observable patterns over time [113].

Fig. 2.1 panel (b) shows the inferred transition and emission probabilities for a 2-state Hidden

Markov Model, including conversations of all lengths. The shown probabilities represent averages over multiple model fits, with standard errors consistently below 0.001. These results provide evidence of two qualitatively different states. Comparing the probability of ending a conversation in each state, $P(X1|Z1)$ and $P(X1|Z2)$, reveals that only $P(X1|Z2) = 0.3$ is nonzero, while $P(X1|Z1) = 0$, indicating that conversations can only end in state $Z2$. This suggests that $Z2$ represents the likely terminal state of a conversation, whereas Z_1 corresponds to earlier stages of interaction.

To further characterize the difference between states Z_1 and Z_2, we analyzed the relative risk of toxic activity (X_3) in state Z_1 compared to Z_2, denoted as RR_{X_3}, and defined as follows:

$$RR_{Z_1} = \frac{P(X_3|Z_2)}{P(X_3|Z_1)} \tag{2.1}$$

A relative risk $RR_{Z_1} > 1$ indicates that toxicity is more likely to occur when a conversation is in state Z_2, whereas a relative risk $RR_{X_1} < 1$ would indicate that toxicity is more probable in state Z_1. Lastly, $RR_{Z_1} = 1$ means that toxicity is equally likely in both states. Since $RR_{Z_1} = \frac{0.08}{0.06} > 1$, the likelihood of encountering toxic content is greater in state Z_2 than in state Z_1. As Z_2 is also the only state where a conversation can end, this finding suggests an association between toxicity and ending a conversation. We evaluated the sensitivity of these results to the substantial presence of short conversations in the dataset. Robustness checks revealed no such sensitivity, as the pattern remained consistent even after progressively filtering out the shortest conversations, from those of length 2 up to a minimum size of 5, and after grouping conversations by channels and the topics they address. See chapter's appendix (see Fig. A.4) for detailed results of this analysis.

While we initially constrained the model to two hidden states, robustness tests with additional states show that a 4-state model fits the data better, see Fig. A.3. Fig. 2.4 shows results for the 4-state model fit, with transition probabilities greater than 0.05 being depicted in panel (a) and the distributions of relative risks of toxic content, RR_{Z_i}, with mean estimates (red dots) and the dashed blue line representing $RR = 1$ shown in panel (b). Transition probabilities show two distinct conversation archetypes. The first archetype involves conversations transitioning among states Z_1, Z_2, and Z_4, while the second consists of conversations that predominantly begin, persist, and conclude in Z_3. Similar to the previously observed 2-state model, there exists a state with a high probability of ending a conversation—namely, Z_4, where $P(X_1|Z_4) = 0.46$. While Z_3 has a nonzero probability of ending a conversation ($P(X_1|Z_3) = 0.02$), this value remains relatively low. In contrast, neither Z_1 nor Z_2 are capable of ending a conversation, as $P(X_1|Z_1) = P(X_1|Z_2) = 0.00$. This suggests that conversations following the first archetype are more dynamic, transitioning through multiple states before potentially ending in Z_4, whereas conversations in the second archetype are more static, largely confined to Z_3.

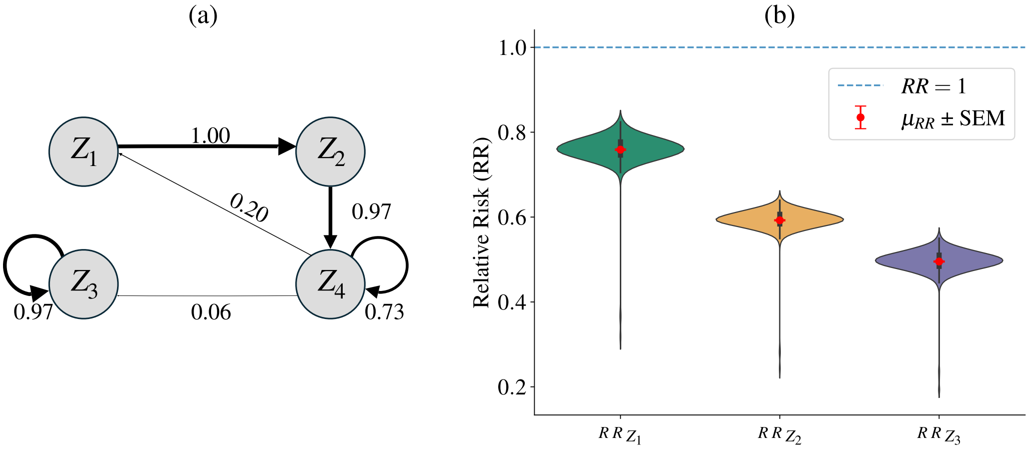


Figure 2.4: **Inferred transition and emission probabilities of a 4-state Hidden Markov Model across all conversations.** **(a)** The inferred transition dynamics of the 4-state Hidden Markov Model, showing only transition probabilities > 0.05. Results reveal two primary conversation patterns: one involving transitions among states Z_1, Z_2, and Z_4, and another where conversations predominantly start, continue, and end in Z_3. Notably, Z_4 has a high probability of ending conversations ($P(X_1|Z_4) = 0.46$). While Z_3 has a small chance of termination ($P(X_1|Z_3) = 0.02$), neither Z_1 nor Z_2 can end a conversation ($P(X_1|Z_1) = P(X_1|Z_2) = 0.00$). This suggests that conversations either transition through multiple states before ending in Z_4 or remain in Z_3 throughout. **(b)** Violin plots show the distribution of inferred relative risk (RR_{Z_i}) values for the probability of observing toxic content (X_3) given states Z_1, Z_2, or Z_3. The RR_{Z_i} is computed relative to state Z_4, meaning the numerator is always the probability of toxic content in Z_4 ($P(X_3|Z_4)$). The dashed blue line represents $RR = 1$, indicating equal likelihood of toxic content across states. Red dots denote mean RR_{Z_i} values (μ_{RR}), with error bars showing the standard error of the mean (SEM). Since $RR_{Z_i} < 1$ for all Z_i, toxic content is more likely in states Z_1, Z_2, and Z_3 than in Z_4.

Fig. 2.4 panel (b) illustrates the distributions for relative risks of toxic content, RR_{Z_i}, which similarly to equation 2.1 is defined as follows:

$$RR_{Z_i} = \frac{P(X_3|Z_4)}{P(X_3|Z_i)} \tag{2.2}$$

The inferred relative risk (RR_{Z_i}) values indicate that toxic content is more likely in states Z_1, Z_2, and Z_3 compared to Z_4, as all RR_{Z_i} values remain below 1. Violin plots in Fig. 2.4 panel (b) show that mean estimates (red dots) are consistently below the dashed blue line at $RR = 1$, confirming this trend. Standard error bars suggest minimal variability across samples. Additional checks show that filtering short conversations does not significantly alter these probabilities, reinforcing the robustness of these findings. See the supplementary material for detailed results of this analysis, Fig. A.4)

Now, we learn model parameters across different video groups, allowing us to capture aggregated conversational behavior. First, we cluster conversations by their channel of origin and learn parameters for each channel (see Fig. 2.5). Secondly, we categorize videos based on their

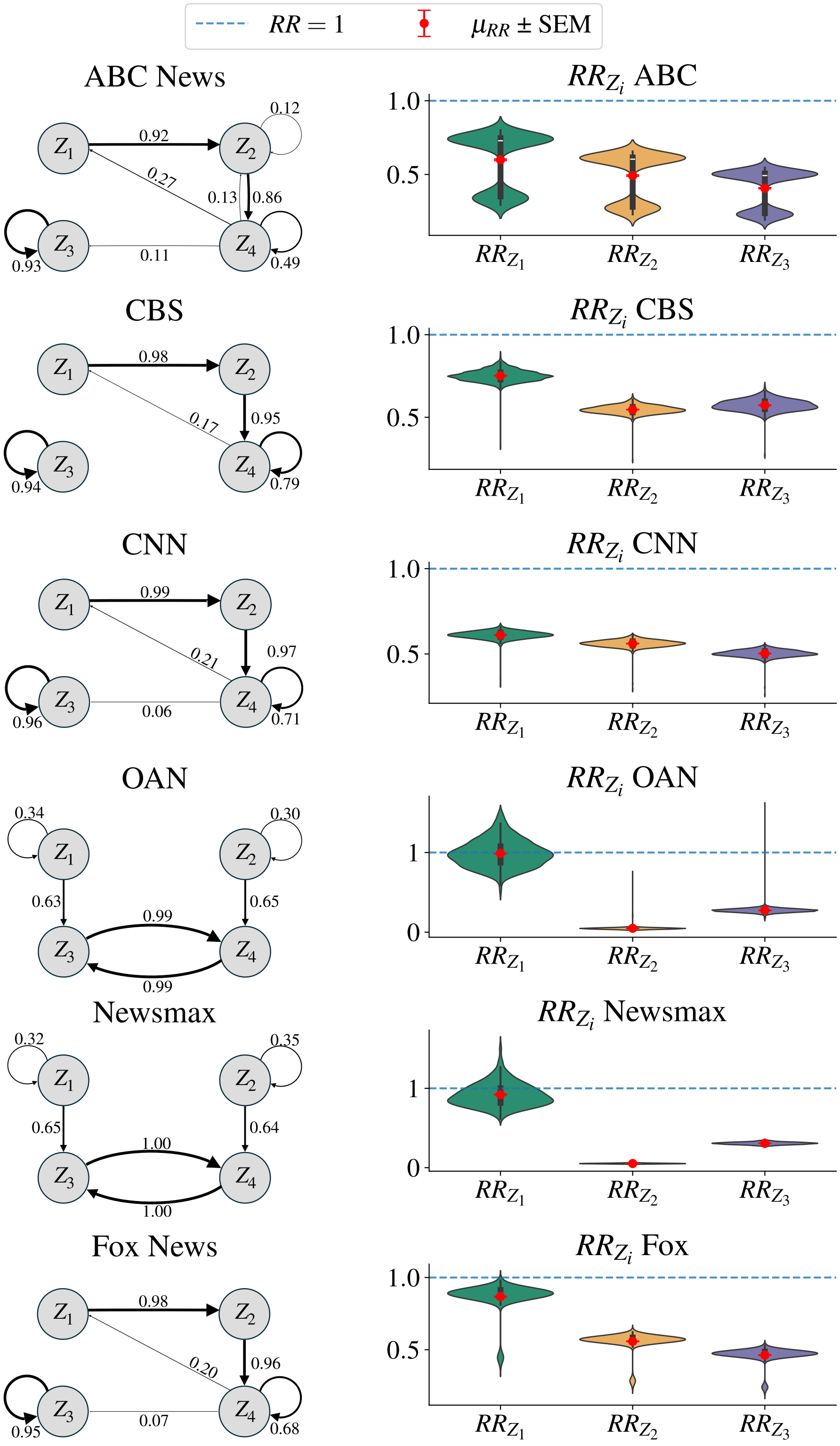


Figure 2.5: **Inferred transition and emission probabilities of a 4-state Hidden Markov Model, grouping conversations by news channel.** The left column displays diagrams of fitted transition probabilities across four latent states (Z_1–Z_4) for each channel. The right column presents violin plots showing the distribution of relative risk (RR_{Z_i}) values for observing toxic content (X_3) in each state, with state Z_4 serving as the reference. The dashed blue line represents $RR = 1$, indicating equal toxicity likelihood across states, while red dots denote mean RR_{Z_i} values with standard error bars ($\mu_{RR} \pm SEM$)

topics, independent of their source channels (see Fig. 2.7).

Fig. 2.5 illustrates the inferred conversational dynamics and toxicity risks across six news channels. The left column presents variations in transition probabilities between latent states, which largely align with the patterns observed in the overall conversation model shown in Fig. 2.4. However, OAN and Newsmax display distinct conversation structures, with two different patterns: one beginning in state Z_1 and transitioning between states Z_3 and Z_4, and another starting in state Z_2 before transitioning to Z_3 and Z_4. These differences may stem from the fact that conversations on OAN and Newsmax tend to be shorter and more uniform in length compared to other channels. Additionally, the relative risk values for all channels remain below 1, indicating that toxic content is less likely to be observed in state Z_4, which is associated with the end of a conversation. This suggests that toxicity is more prevalent in any other conversational state Z_1, Z_2, or Z_3, reinforcing the idea that toxic interactions are more likely to occur during ongoing discussions rather than at their conclusion. Lastly, the bimodal distribution of relative risk values for ABC suggests the presence of at least two distinct patterns for toxic content within this channel.

We categorize videos based on their topics, independent of their channel sources. By learning parameters for video groups, our model produces results that summarize the aggregated behavior of video ensembles. In the previous section, we grouped videos by their news media channel publishers and observed that negative sentiment tends is more likely to arise toward the end of conversations. To ensure that the observed connection between negative sentiment and disengagement is not dependent on group definitions, we propose an alternative way to define video ensembles. This approach leverages topic modeling, a scalable machine-learning technique that organizes and classifies text by identifying semantically related content [125].

We use a hierarchical stochastic block model (hSBM) [126] approach to cluster 19,365 video descriptions and 26,041 words, creating a network comprising 576,092 edges. The model automatically identifies the number of topics and hierarchical levels, eliminating the need for prior specification. Figure 2.6 demonstrates the application of the hierarchical Stochastic Block Model (hSBM) on video descriptions. The clusters on the right side represent inferred topics, while the left side correspond to video groupings. At the highest level, the model separates word nodes from video description nodes, reflecting its bipartite structure. At the fourth level of hierarchy, it categorizes words into 10 topics, with two topics (0 and 4) representing functional words. (The model's ability to recognize function words provides a data-driven alternative to the traditional practice of manually removing stopwords.)

Unlike conventional models such as LDA, hSBM automatically clusters documents. This feature allows us to group video descriptions based on the topics they address. Additionally, hSBM offers a framework that leverages the similarities between topic modeling and community detection in complex networks and reinterprets topic modeling as a community detection

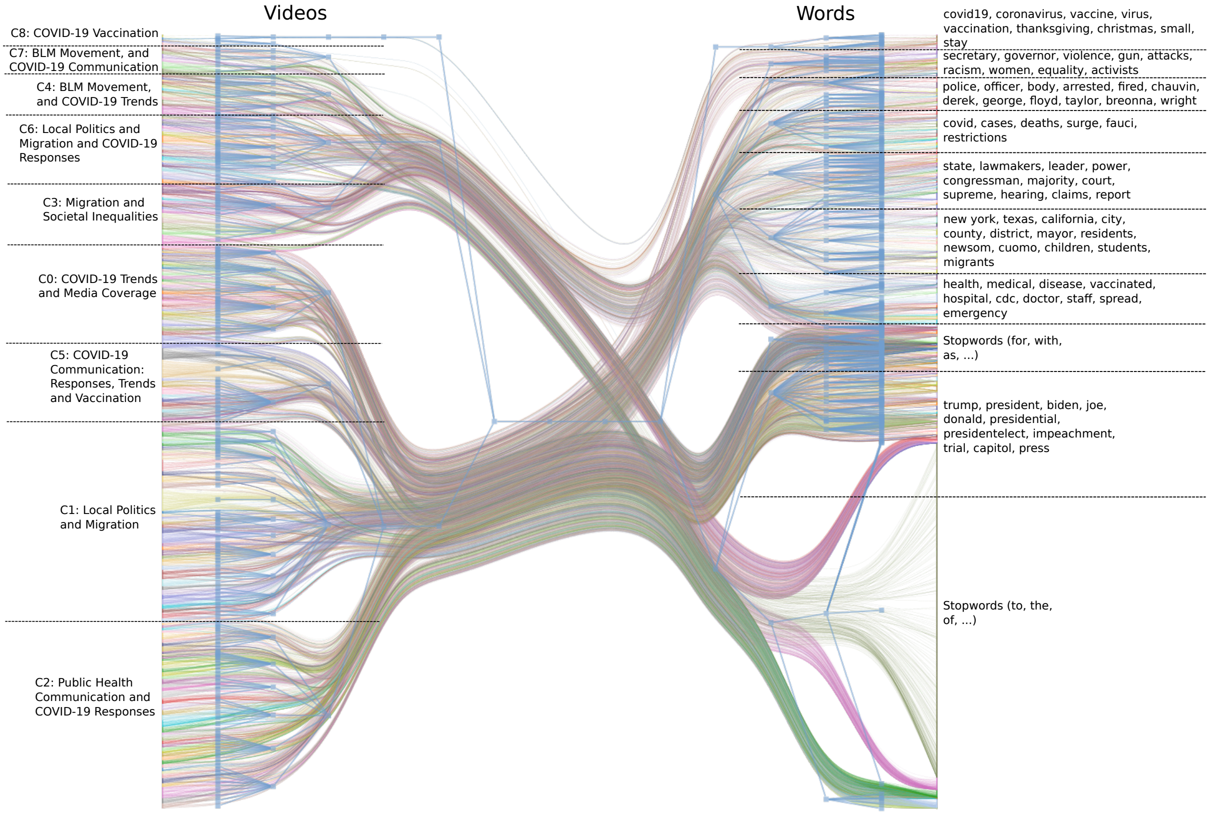

Videos
Words
C8: COVID-19 Vaccination
C7: BLM Movement, and COVID-19 Communication
C4: BLM Movement, and COVID-19 Trends
C6: Local Politics and Migration and COVID-19 Responses
C3: Migration and Societal Inequalities
C0: COVID-19 Trends and Media Coverage
C5: COVID-19 Communication: Responses, Trends and Vaccination
C1: Local Politics and Migration
C2: Public Health Communication and COVID-19 Responses
covid19, coronavirus, vaccine, virus, vaccination, thanksgiving, christmas, small, stay
secretary, governor, violence, gun, attacks, racism, women, equality, activists
police, officer, body, arrested, fired, chauvin, derek, george, floyd, taylor, breonna, wright
covid, cases, deaths, surge, fauci, restrictions
state, lawmakers, leader, power, congressman, majority, court, supreme, hearing, claims, report
new york, texas, california, city, county, district, mayor, residents, newsom, cuomo, children, students, migrants
health, medical, disease, vaccinated, hospital, cdc, doctor, staff, spread, emergency
Stopwords (for, with, as, ...)
trump, president, biden, joe, donald, presidential, presidentelect, impeachment, trial, capitol, press
Stopwords (to, the, of, ...)

Figure 2.6: **Hierarchical stochastic block model (hSBM) fit to video documents.** Results of clustering 18,627 video titles and descriptions, and 26,041 words, forming a network with 576,092 edges to infer topics and video clusters using the hSBM approach. At the highest hierarchical level, the model separates word nodes from video description nodes, reflecting the network's bipartite structure. At the fourth hierarchical level, the model categorizes words into 10 topics, with two topics representing functional words, shown on the left-hand side of the bipartite matrix, and identifies 9 distinct video clusters, shown on the right-hand side of the bipartite matrix.

problem by representing the word-document matrix as a bipartite network. By utilizing community detection techniques for topic modeling, the hSBM approach constructs a nonparametric Bayesian model grounded in a hierarchical stochastic block model (hSBM), successfully overcoming many of LDA's limitations [126].

Fig. 2.6 presents a bipartite network representation of our corpus, along with a hierarchical stochastic block model fit to the video documents to infer topics and identify video clusters. The clusters on the right represent inferred topics, while the clusters on the left correspond to video clusters. At the highest level, the model separates word nodes from video description nodes, reflecting its bipartite structure. For this analysis, we focus on the fourth hierarchical level, where the model categorizes words into 10 topics, including two functional word topics, and identifies nine video clusters. Following the approach outlined in the previous section, we used these clusters to fit a 4-state Hidden Markov Model to each video cluster, as illustrated in

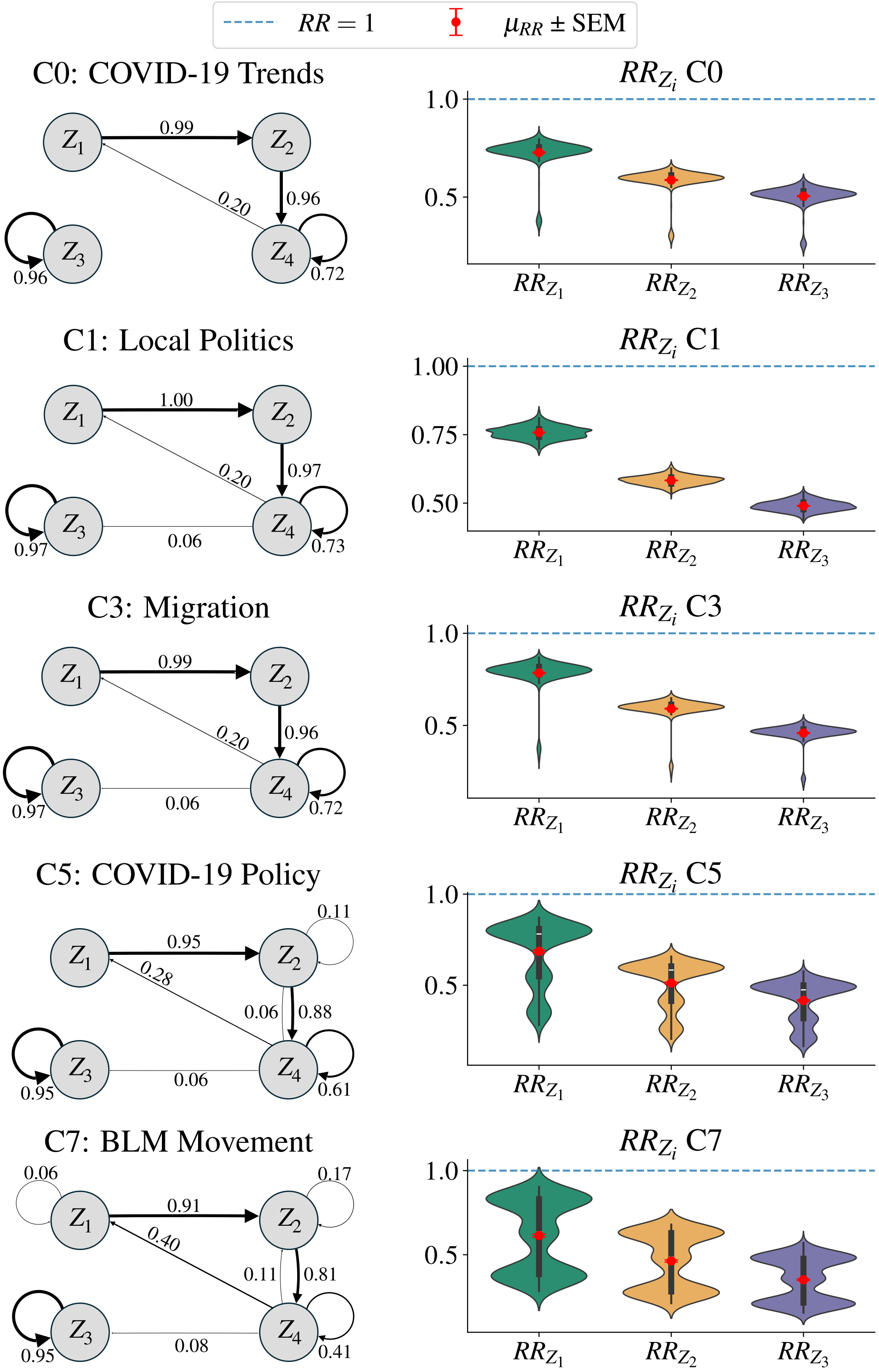


Figure 2.7: **Inferred transition and emission probabilities of a 4-state Hidden Markov Model, grouping conversations by topic clusters** The left column displays diagrams of fitted transition probabilities across four latent states (Z_1–Z_4) for featured cluster. The right column presents violin plots showing the distribution of relative risk (RR_{Z_i}) values for observing toxic content (X_3) in each state, with state Z_4 serving as the reference. The dashed blue line represents $RR = 1$, indicating equal toxicity likelihood across states, while red dots denote mean RR_{Z_i} values with standard error bars ($\mu_{RR} \pm SEM$)

Figure 2.1. Our primary focus continues to be on the emission probabilities, which enable us to characterize and compare the hidden states effectively.

Fig. 2.7 presents results for selected video clusters. Similar to the channel cluster results in Fig. 2.5, the transition probabilities between latent states closely follow the patterns observed in the overall conversation model depicted in Fig. 2.4. As with ABC News, two video clusters—Cluster 5 (COVID-19 Policy) and Cluster 7 (Black Lives Matter movement)—exhibit a multimodal distribution of relative risk values, suggesting the presence of more than one distinct pattern for toxic content within these topics. Moreover, the distribution of relative risk values indicates systematic differences in how toxicity emerges and persists across conversational states, reinforcing the idea that toxic exchanges are more likely at earlier stages rather than toward the end of conversations. Overall, exploring an alternative categorization of videos reveals similar patterns across different grouping methods, further confirming the robustness of our findings.

2.4 Conclusions

In this study, we analyze both individual comments and the broader conversations they form, defining a conversation as a time-bound sequence of replies to a single top-level comment. We include replies posted within 10 days of the original comment, provided they directly respond to it, even if they mention other users. This approach yields a dataset of over 2.9 million conversations. Our findings suggest a connection between heightened political moments and increased online hostility, as well as an influence of initial sentiment on the overall tone of discussions. Using Hidden Markov Models [113–115], we observe that while a 2-state model indicates an association between toxicity and conversation termination, this link diminishes in models with more states. A more refined (and better fitting) 4-state model, which captures more detailed latent states, further reveals that toxicity is more prevalent in the early stages of a conversation. These insights contribute to a deeper understanding of online political discourse and underscore the need for a more nuanced perspective when examining the role of toxicity in shaping digital interactions.

Political segregation and the lack of conversation across ideological divides have become increasingly prevalent phenomena, posing significant risks to democratic discourse and the institutions underpinning it [4, 19]. Social media platforms, as key players in modern communication, occupy a central role in this process [127]. These platforms, driven by algorithms designed to maximize engagement and monetary incentives, often create environments where political ideology clashes with marketing goals and user interactions [28]. A key outcome of such environments is the proliferation of antisocial behavior, such as toxicity and insults, primarily driven by small but vocal minorities. These behaviors significantly influence broader participation, fostering disengagement and self-censorship among the majority.

Our analysis of YouTube during the 2020 US presidential elections reveals several important findings. First, toxic and insulting behaviors are prevalent in online political conversations, particularly during politically charged events. We observed significant spikes in toxicity and insults coinciding with major political moments such as the Black Lives Matter protests, Election Day, and the Capitol riots of January 6. Furthermore, toxic top-level comments are more likely to elicit similarly negative replies, suggesting a cyclical relationship where toxicity fuels further toxicity. These patterns illustrate how toxicity is not an isolated phenomenon but an entrenched feature of political discourse online, particularly during moments of increased political tension. Together, these findings indicate that toxicity in online spaces reflects societal instability and is a driver of further polarization and disengagement.

The implications of these findings are far-reaching for understanding the dynamics of online political discourse. The spikes in toxicity observed during politically significant events suggest that online discourse is highly responsive to real-world political developments. This responsiveness reflects the emotional and ideological intensity of these moments, with online platforms serving as spaces for both expression and conflict [121]. At the same time, the observation that toxicity generates further toxicity highlights how these platforms may reinforce and amplify negative interactions.

The cyclical nature of toxicity is particularly concerning because it can alienate a large portion of the user base. When individuals encounter toxic environments, they may disengage, self-censor, or even abandon the platform entirely. This dynamic reduces the diversity of perspectives in online conversations, skewing public discourse and misrepresenting public opinion. As such, online toxicity reflects polarization and actively contributes to it, undermining the potential of digital spaces to serve as arenas for democratic deliberation [80, 128].

While this study provides meaningful insights, it is important to acknowledge its limitations. First, our analysis focuses on specific YouTube communities belonging to US news outlets, which, while providing detailed insights into conversations surrounding key political events in that country, narrows the scope of our results. The choice to study particular communities was intentional, allowing us to delve deeply into relevant discussions. However, the findings may not fully represent broader online discourse or apply to different contexts [70, 76, 129].

Although YouTube hosts a vast number of channels—many of which promote politically relevant content such as far-right ideologies, conspiracy theories, and ”anti-woke” material—the scale and diversity of this content are beyond the scope of this project [78]. Capturing a more comprehensive picture of all politically relevant conversations on YouTube would require datasets that include a broader range of independent creators and communities. This limitation suggests that future research should aim to expand datasets and examine a wider array of political content and interactions across different platforms.

Additionally, the influence of algorithmic recommendation systems and content moderation technologies adds complexity to the relationship between toxicity and self-censorship [79]. The role of these technologies in shaping user behavior and interactions remains an open question. Comparative analyses across platforms and sociopolitical contexts could clarify whether these dynamics are unique to YouTube or generalizable to other algorithm-driven spaces. A recent example of cross-platform analysis is the work by Avalle et al. [80], where authors found that toxicity and user participation in debates operate independently.

We partially confirm these results; however, our study differs in one critical aspect: the definition of a conversation. As a result, this difference in definition makes direct comparisons between our findings and prior results challenging. We define a conversation as a coherent, time-bound exchange of posts focused on a single top-level comment and its replies. This approach ensures that we capture user interactions centered on a specific topic. On the other hand, Avalle et al. [80] define a conversation as an entire thread of a video's comments, ordered chronologically. We believe this approach does not ensure that users interact with the previous comment in the sequence, as implied by the definition of a conversation, for two key reasons. First, YouTube does not automatically sort comments chronologically, making it highly likely that users are not engaging with comments in this order. Second, there are instances where significant time lapses occur between comments. In such cases, users' posts may be influenced more by the content of the video itself than by the most recent comment.

The relationship between online toxicity, self-censorship, and disengagement has profound implications for how we interpret ideological polarization and public discourse in digital spaces. Public discourse online may not accurately reflect individuals' private beliefs or offline behaviors [62, 65, 66]. This discrepancy risks distorting our understanding of public opinion and intensifying polarization.

Nevertheless, as the boundaries between online and offline activities blur, especially in political discourse, understanding the role of social media platforms, recommendation algorithms, and collective antisocial behaviors becomes increasingly urgent. Social media platforms accelerate the dissemination of information, often reinforcing feedback loops that amplify toxicity and polarization. Our findings suggest a circular causality between toxicity and disengagement, where negative behaviors discourage meaningful participation, further entrenching polarization and alienation.

Addressing these dynamics is critical for preserving democratic engagement in a digital age. Platforms must consider how their design and algorithms influence discourse, fostering environments that support diverse perspectives and constructive dialogue. As online spaces increasingly shape public opinion and democratic processes, understanding and mitigating the effects of toxicity are essential steps toward ensuring their positive contribution to society.

Chapter 3

Teams work better at evaluating and annotating misleading political content

3.1 Introduction

Social media platforms face growing scrutiny for their role in the widespread dissemination of misinformation. To address this issue, platforms have implemented various strategies, including community-based content moderation, as an alternative to expert labeling and potentially biased machine-learning algorithms. Although community-based moderation has received positive feedback, its effectiveness in combating misinformation remains uncertain.

This study evaluates X's Community Notes feature, which enables users to label tweets as misleading or accurate and provide contextual notes. Users can also rate the helpfulness of these labels and notes [130]. Previous research has found that ”Birdwatchers” (community members contributing to Community Notes) often focus on counter-partisan content [131]. However, one notable limitation of this system is the absence of collaborative note creation. We hypothesize that enabling user collaboration in labeling and annotating content would yield more neutral and accurate outcomes.

To test this hypothesis, we conducted an online experiment where participants collaborated on content moderation tasks. The study examines how group composition and the visibility of party affiliation influence the accuracy of fact-checking. In the experiment, participants were assigned the task of composing individual and collaborative notes for 40 political tweets, sourced from Democrat and Republican accounts. Participants–recruited via Prolific–were randomly paired with partners from either the same or different political affiliation, resulting in three group configurations: (1) Democrat-Democrat (DD), (2) Republican-Republican (RR), and (3) Democrat-Republican (DR). Moreover, each team was randomly assigned to one of two treatments: in Treatment 1 (the ‘Overt’ treatment), participants could view each other's political affiliations, while in Treatment 2 (the ‘Covert’ treatment), participants were unaware of each other’s affiliations.

Community-based solutions employ a collaborative approach to problem-solving. As opposed to a crowd-sourced solution that leverages the wisdom of the crowd approach, community-based solutions harness collective intelligence (CI), that is, the ability of a group to solve complex problems and make decisions collaboratively [101–103, 132]. Research indicates that collective intelligence has the potential to enhance political trust and promote well-informed, inclusive political discourse [133, 134]. This forms the basis of our first hypothesis:

> **H1:** Teams outperform individuals in evaluating and annotating misleading political content

A community-based verification system may encounter pitfalls and shortcomings if it overlooks group dynamics. Diverse groups excel in collaborative tasks like Wikipedia edits and rating news headlines [98, 135–138]. They also contribute to reducing polarization in discussions, particularly, especially when partisan identities remain undisclosed [139]. However, when party affiliation is visible, individuals tend to favour information aligned with their own party and discredit opposing viewpoints [140]. For instance, research by Allen et al. on X’s Community Notes system reveals users’ tendency to offer negative evaluations of tweets from opposing parties and perceive their evaluations as unhelpful [131]. This behaviour may stem from individuals associating the discussion with familiar partisan stereotypes, past experiences, and personal beliefs [140]. This forms the basis of our second set of hypotheses:

> **H2a:** Heterogeneous groups outperform homogeneous groups in evaluating and annotating misleading political content.

> **H2b:** When party affiliation is known, homogeneous groups outperform heterogeneous groups in evaluating and annotating misleading political content.

3.2 Data and methods

3.2.1 Experimental design

On January 23, 2021, X launched Birdwatch, now known as Community Notes, initiating a community-driven effort to moderate misleading content. Community Notes aimed to harness X’s diverse user base for content moderation, presenting selected tweets to Birdwatchers who were tasked with contextualizing them and identifying whether they were misleading, providing reasoning for their judgment. Other Birdwatchers would then rate the helpfulness of these notes. A note is only attached to the original tweet and displayed to all users if it accumulates sufficient helpful ratings. Originally, the helpfulness score was determined by the difference between positive and negative ratings. However, this method was revised because Birdwatchers were observed to label tweets and rate notes in a highly polarized manner [131, 141]. Following the adjustment, a note was considered of higher quality if it received helpful ratings from Birdwatchers with diverse rating patterns [130].

While Community Notes benefits from the wisdom of the crowds, we argue that its effectiveness could be enhanced by integrating Collective Intelligence. Research indicates that diverse groups achieve optimal performance by interacting and merging personal and social information [142]. In our study, we conducted an experiment mirroring the Community Notes process, where users were paired and instructed to collaborate on writing a note they both endorsed.

In the summer of 2023, we enlisted 480 active Prolific participants who self-identified as either Democrats or Republicans. They were randomly paired with partners from either the same or different political affiliations, resulting in three group configurations: (1) Democrat-Democrat (DD), (2) Republican-Republican (RR), and (3) Democrat-Republican (DR). Each team was randomly assigned to one of two treatments: in Treatment 1 (the 'Overt' treatment), participants could see each other's political affiliations, while in Treatment 2 (the 'Covert' treatment), participants were unaware of each other's affiliations.

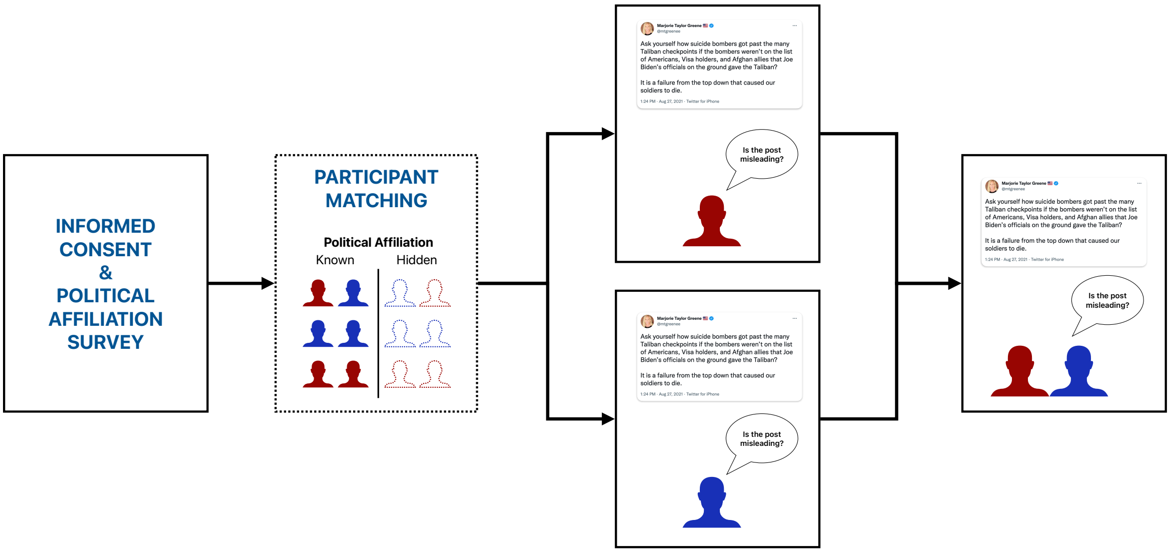
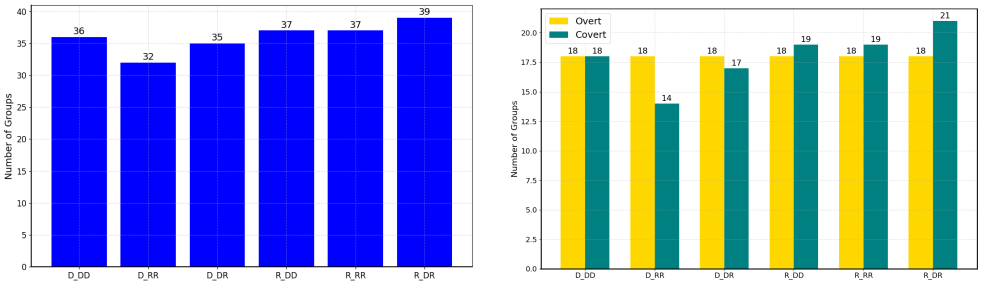

INFORMED CONSENT & POLITICAL AFFILIATION SURVEY
PARTICIPANT MATCHING
Political Affiliation
Known
Hidden
Marjorie Taylor Greene
@mtgreenee
Ask yourself how suicide bombers got past the many Taliban checkpoints if the bombers weren't on the list of Americans, Visa holders, and Afghan allies that Joe Biden's officials on the ground gave the Taliban?
It is a failure from the top down that caused our soldiers to die.
1:24 PM · Aug 27, 2021 · Twitter for iPhone
Is the post misleading?

Figure 3.1: **Collaborative annotation experimental design.** Participants wrote individual and collaborative notes for 40 political tweets, sourced from Democrat and Republican accounts. They were randomly paired with partners from either the same or different political affiliations, resulting in three groups: Democrat-Democrat (DD), Republican-Republican (RR), and Democrat-Republican (DR). Each team was then randomly assigned to one of two treatments: 'Overt' treatment, where participants could see each other's political affiliations, and 'Covert' treatment, where participants were unaware of each other's affiliations.

Once paired, participants assessed the credibility of one of 40 tweets, sourced from both Democratic and Republican accounts. Initially, they evaluated the tweets individually, then repeated the process with their partner (see Figure 3.1). Their evaluation involved determining whether, based on the latest available evidence, the tweet was misinformed or potentially misleading. Additionally, they explained their reasoning in writing within a 280-character limit by providing context that they felt would help others understand why they believe the tweet to be misleading or not. Participants were encouraged to cite external sources. Collaboration was a pre-requisite for successful completion of the study, thus, participants were instructed to engage in conversation with their partner using the chat box. In the ‘Overt’ treatment, chat nicknames reflected political affiliation (e.g., ‘Democrat 1’, ‘Republican 2’), whereas in the ‘Covert’ treatment, handles were generic (‘Participant 1’, ‘Participant 2’). The participants might have been able to guess their partner’s political affiliation by chatting with them and discussing their views on the tweet, however, this process was covert. Finally, participants completed a brief survey to provide insights into their labeling rationale and overall experience.



(a) (b)

Figure 3.2: **Breakdown of total number of collaborative notes** ($N = 238$) **by group configuration and treatments** We aimed for each of the 6 treatment groups to assess each tweet at least once, targeting a minimum of 20 evaluations per category. Reported numbers exceeding 20 may stem from uncertainties in the recruitment process. Notes failing to meet quality standards were excluded, leading to reported numbers below the 20-evaluation target.

All in all, our analysis includes 648 notes, consisting of both collaborative and individual evaluations (216 each). Figure 3.2 illustrates the breakdown of collaborative notes (216) by group configuration and treatments. Our goal was to assess each tweet at least once by each of the 6 treatment groups, aiming for a minimum of 20 successful evaluations per category. Occasionally, the reported number exceeds 20 due to uncertainties in the recruiting process. This is due to the fact that Prolific has more active participants who self-identified as Democrats (∼10,000) than Republicans (∼3,000). Therefore, to optimize waiting times we had to occasionally pair a participant with whoever was active and available. Likewise, there were instances where we fell short of the 20-evaluation target. Following the experiment, we meticulously reviewed the data to ensure minimum quality standards were met. We discarded interactions where either participant ceased communication or instances where chat text was submitted as a note. Any note where participants failed to reach an agreement or comprehend the task were discarded.

Finally, to mirror the Community Notes helpfulness rating system, we recruited 1610 additional Prolific users, evenly divided between self-identified Democrats and Republicans (805 each), who had not taken part in the Collaborative Study. Their task was to evaluate the notes and rate their helpfulness on a scale from 1 to 10. In addition to the crowd-sourcing ratings, we sought evaluations from three experts on the same set of notes. In the following section, we use both the crowd-sourced ratings and the expert assessments as indicators of notes' helpfulness.

3.2.2 Helpfulness and improvement measures

On average, each note received ratings from 5 Democrats and 5 Republicans. The average of the ratings from Democrats on note a will be referred to as the helpfulness score from Democrats (H^a_{D}), while the average of the ratings from Republicans on note a will be referred to as the helpfulness score from Republicans (H^a_{R}). Average expert ratings on note a will be referred to as the helpfulness score from experts (H^a_{E}). The helpfulness scores for on note a are calculated using the following equations:

$$H^a_{\mathrm{D}} = \frac{1}{n_{\mathrm{D}}} \sum_{i=1}^{n_{\mathrm{D}}} h^a_{D_i}, \quad H^a_{\mathrm{R}} = \frac{1}{n_{\mathrm{R}}} \sum_{i=1}^{n_{\mathrm{R}}} h^a_{R_i}, \quad H^a_{\mathrm{E}} = \frac{1}{n_{\mathrm{E}}} \sum_{i=1}^{n_{\mathrm{E}}} h^a_{E_i}$$

where n_{D}, n_{R}, and n_{E} refer to the number of Democrats, Republicans, and experts that rated note a, respectively, and $h^a_{D_i}$, $h^a_{R_i}$, and $h^a_{D_i}$ refer to the evaluation by user i on note a from Democrats, Republicans, and experts, respectively.

To calculate the potential improvement of the notes written by the teams, new variables called I_{D}, I_{R}, and I_{E} are introduced, representing the improvement in helpfulness as rated by Democrats, Republicans, and experts, respectively. The improvements are calculated using the following equations:

$$\begin{aligned} I_{\mathrm{D}} &= H^{ab}_{\mathrm{D}} - \frac{1}{2}\left(H^a_{\mathrm{D}} + H^b_{\mathrm{D}}\right), \\ I_{\mathrm{R}} &= H^{ab}_{\mathrm{R}} - \frac{1}{2}\left(H^a_{\mathrm{R}} + H^b_{\mathrm{R}}\right), \\ I_{\mathrm{E}} &= H^{ab}_{\mathrm{E}} - \frac{1}{2}\left(H^a_{\mathrm{E}} + H^b_{\mathrm{E}}\right), \end{aligned}$$

where H^{ab}_x refers to the helpfulness rating of the note written by the team that authored notes a and b individually (with x replaced by D, R, or E for Democrats, Republicans, and experts, respectively).

Table 3.1: **Collaboration improves the notes' helpfulness.** Results from two-sample hypothesis tests the helpfulness of individually- and collaboratively-written notes. Expert ratings indicate that collaboratively written notes are more helpful than individual notes in contextualizing the original Tweets, with significance at the 1% confidence level. Helpfulness scores from Republicans also support this finding at the 5% confidence level

Category	**Individuals** $\mu(\sigma)$	**Teams** $\mu(\sigma)$	**T-value**	***p*-value** **(parametric)**	***p*-value**[1] **(non-parametric)**
H_E	2.9 (2.0)	3.4 (1.9)	-2.8	0.0049**	0.0018**
H_D	4.7 (1.9)	4.9 (1.7)	-1.3	0.1844	0.0926
H_R	4.6 (1.5)	4.8 (1.5)	-2.0	0.0456*	0.0240*

1. Results are based on two-sample hypothesis tests using the Bootstrap method

(*) p<0.05; (**) p<0.01; (***) p<0.001

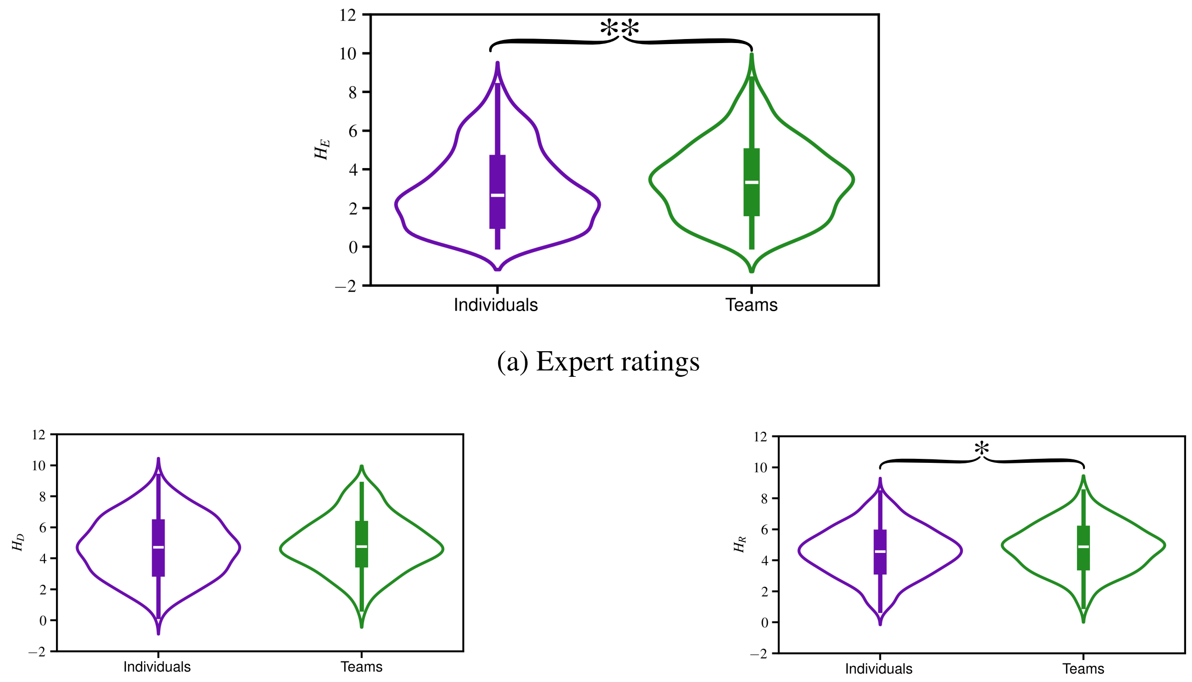
3.3 Results

3.3.1 Collaboration improves annotations

Our initial hypothesis suggests that notes produced by teams will be rated more helpful than those authored individually. To test this, we compare the average helpfulness scores for the notes written by individuals before participating in the collaborative tasks to the scores of notes written collaboratively by teams.

Figure 3.3 illustrates the distribution of helpfulness scores both from crowd-sourced and expert assessments. Helpfulness ratings from experts are represented by H_{E} and shown in Figure 3.3a, while ratings from self-identified Democrats and Republicans are denoted as H_{D} and H_{R}, and shown in Figure 3.3b and Figure 3.3c, respectively. Table 3.1 presents the results of two-sample hypothesis tests comparing individual and team-authored notes, incorporating both crowd-sourced and expert assessments.

Expert ratings show that the collaboratively written notes are more helpful than individually written notes in contextualizing the original Tweets. This results is significant at the 1% confidence level. Helpfulness scores from Republicans support this finding, showing a significance at the 5% confidence level. However, Democrats' helpfulness scores reveal no significant differences between individual and collaborative notes. It is important to exercise caution when interpreting the ratings from Republicans and Democrats, as these scores may be influenced by inherent biases associated with political affiliation [131].



(a) Expert ratings

(b) Democrats' ratings

(c) Republicans' ratings

Figure 3.3: **Distribution of helpfulness scores both from crowd-sourced and expert assessments.** Expert ratings, H_{E}, are shown in panel (a), while ratings from self-identified Democrats and Republicans, H_{D} and H_{R}, are shown in panels (b) and (c), respectively. Expert ratings indicate that collaboratively written notes are more helpful than individual notes in contextualizing the original Tweets, with significance at the 1% confidence level (denoted by **). Helpfulness scores from Republicans also support this finding at the 5% confidence level (denoted by *).

3.3.2 Gains from collaboration depends on group's political composition

According to our second hypothesis, we predict that notes authored by diverse groups will, overall, receive higher helpfulness ratings compared to those written by homogeneous groups. Additionally, when party affiliation is known, we anticipate homogeneous groups to produce more helpful notes. To assess these hypotheses, we analyze the average helpfulness scores across various groups, including the source of the evaluated tweet and the treatment. We distinguish between crowd-sourced and expert assessments in our analysis.

Figure 3.4 shows the distribution of notes' improvement indexes from experts (I_E) by group's political composition, while Table 3.2 presents the results of two-sample hypothesis tests comparing these indexes across groups. The reference category is the homophilic team, aligned with a tweet's political leaning — RR for Republican tweets and DD for Democrat tweets. For Republican tweets, treatment groups include mixed teams (DR) and heterophilic teams (DD). For Democrat tweets, the reference category remains the homophilic team (DD), with mixed teams (DR) and heterophilic teams (RR) as treatment groups.

For evaluations of Democrat tweets, neither heterophilic nor mixed teams outperform the homophilic team, that is, the performance across all three team compositions is comparable. However, for evaluations of Republican tweets, heterophilic teams (DD) perform comparably to homophilic teams (RR). Notably, mixed teams (DR) produce better notes than homophilic teams (RR) — a result that is statistically significant at the 5% confidence level for both parametric and non-parametric tests.

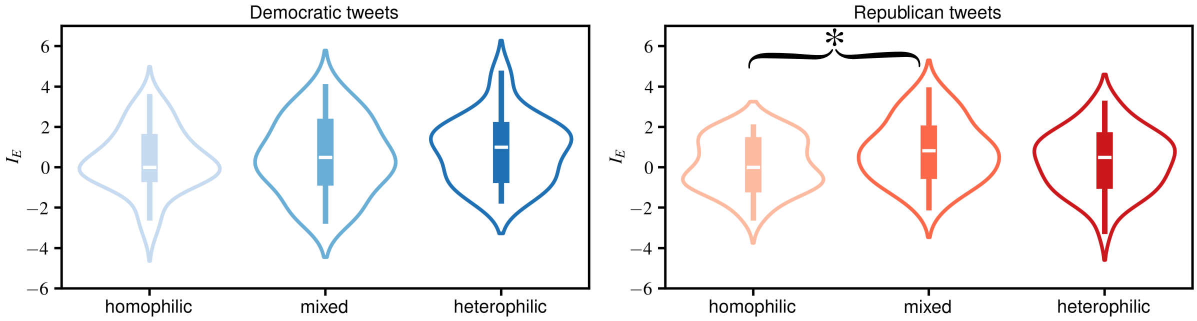


Figure 3.4: **Distribution of notes' improvement indexes by group's political composition.** The left panel shows evaluations of tweets from Democratic sources, with homophilic (DD), mixed (DR), and heterophilic (RR) teams. The right panel shows evaluations of tweets from Republican sources, with homophilic (RR), mixed (DR), and heterophilic (DD) teams. Mixed teams (DR) produce significantly better notes than homophilic teams (RR) for Republican tweets (5% significance level, indicated by $*$), while no team outperforms the homophilic group for Democrat tweets.

Table 3.2: **Gains from collaboration by group's political composition.** Results from two-sample hypothesis tests comparing notes' improvement indexes by experts (I_E) across groups with different political compositions. The reference category is the homophilic team—RR for Republican tweets and DD for Democrat tweets. Treatment groups are mixed teams (DR) and heterophilic teams (DD for Republican tweets, RR for Democrat tweets). Mixed teams (DR) produce significantly better notes than homophilic teams (RR) for Republican tweets (5% significance level), while no group outperforms the homophilic team for Democrat tweets.

Tweet	**Control**	**Treatment**	N_c	N_t	$\mu_c(\sigma_c)$	$\mu_t(\sigma_t)$	**T-value**	***p*-value (parametric)**	***p*-value[1] (non-parametric)**
Republican	RR	DR	35	37	0.11 (1.3)	0.78 (1.5)	-2.0	0.0463*	0.0184*
	RR	DD	35	37	0.11 (1.3)	0.32 (1.5)	-0.6	0.5206	0.2517
Democrat	DD	DR	36	35	0.36 (1.5)	0.5 (1.8)	-0.3	0.7275	0.3596
	DD	RR	36	32	0.36 (1.5)	0.86 (1.6)	-1.3	0.1845	0.0845

1. Results are based on two-sample hypothesis tests using the Bootstrap method

(*) p<0.05; (**) p<0.01; (***) p<0.001

3.3.3 Overt signaling hinders collaborative synergy

To investigate the effect of overt political affiliation signaling on team interactions, we partition the results into two treatment groups: those in which political affiliation was explicitly disclosed and those in which it remained covert.

Figure 3.5 illustrates the distribution shows the comparative distribution of improvement indices for collaborative notes across Covert and Overt treatments, based on crowd-sourced and expert assessments. Figure 3.5a illustrate improvement indices based on expert ratings, I_E, while improvement indices based on ratings from self-identified Democrats and Republicans, I_D and I_R, are shown in Figures 3.5b and 3.5c, respectively. Table 3.3 presents the results of parametric and non-parametric two-sample hypothesis tests comparing improvement indices between Covert and Overt treatments.

Expert ratings indicate that the collaborative advantage in note helpfulness declines when political affiliation is overtly disclosed. This outcome is reinforced by Republican and Democrat ratings, with all findings achieving statistical significance at the 1% level. Furthermore, in some group compositions, overt signaling of political affiliation results in a small negative improvement index, suggesting that collaboratively written notes may be less helpful than individually written notes by the same participants.

Table 3.3: **Overt signaling hinders collaborative synergy.** Results from two-sample hypothesis tests comparing notes' improvement indices between Covert and Overt treatments. Expert ratings on collaborative notes indicate when political affiliation is overtly signaled, the improvement in helpfulness diminishes. Ratings from both Republicans and Democrats support this finding, with all results showing significance at the 1% confidence level.

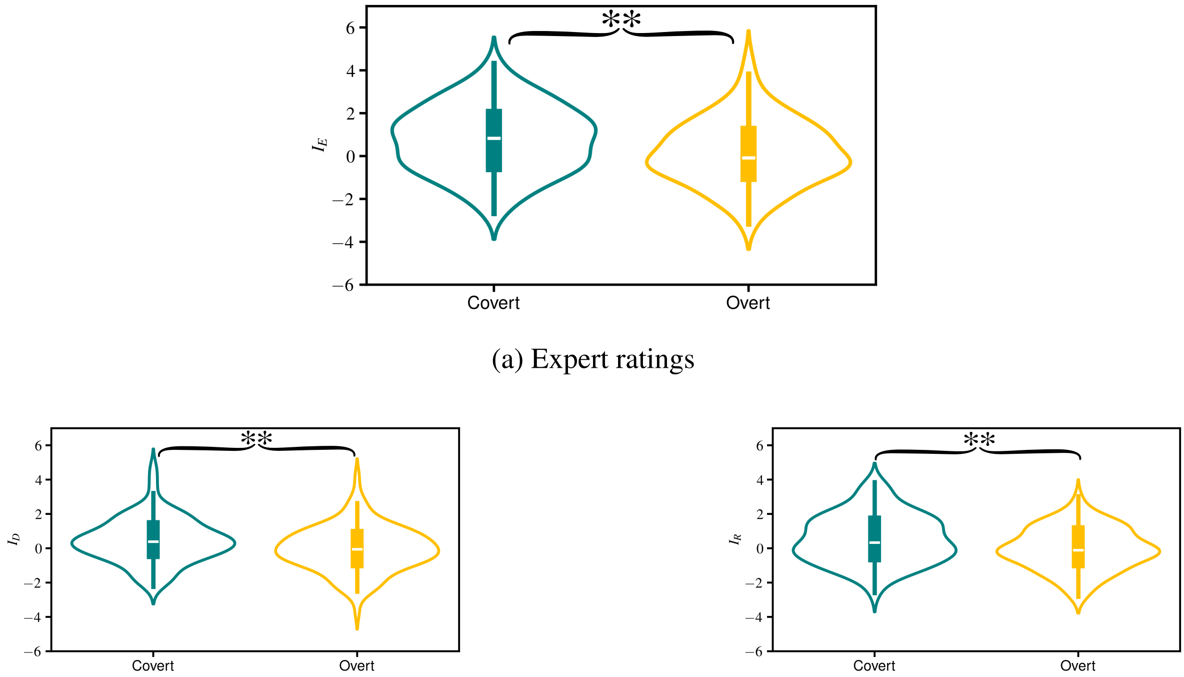
Category	**Covert** $\mu(\sigma)$	**Overt** $\mu(\sigma)$	**T-value**	p**-value (parametric)**	p**-value**[1] **(non-parametric)**
I_E	0.76 (1.5)	0.16 (1.5)	2.9	0.0043**	0.0021**
I_D	0.45 (1.3)	-0.045 (1.3)	2.8	0.0055**	0.0023**
I_R	0.54 (1.4)	-0.037 (1.2)	3.2	0.0014**	0.0003***

1. Results are based on two-sample hypothesis tests using the Bootstrap method

(*) p<0.05; (**) p<0.01; (***) p<0.001

3.4 Discussion and conclusions

Despite the initial acclaim received by community-based fact-checking methods, uncertainties persist regarding their ability to impartially verify facts, particularly in the realm of political misinformation, disinformation, and propaganda [131]. This study aims to address this gap by



(a) Expert ratings

(b) Democrats' ratings

(c) Republicans' ratings

Figure 3.5: **Distribution of improvement indices for collaborative notes under Covert and Overt treatments, based on both crowd-sourced and expert assessments.** Improvement indices based on expert ratings, I_E, are shown in panel (a), while improvement indices based on ratings from self-identified Democrats and Republicans, I_D and I_R, are shown in panels (b) and (c), respectively. Expert ratings suggest that the positive effect of collaboration on note helpfulness weakens when political affiliation is made explicit. Both Republican and Democrat ratings align with this result, each reaching 1% significance (denoted by **).

Table 3.4: **Overt signaling hinders collaborative synergy for Democrat-Democrat teams** Results from two-sample hypothesis tests comparing note improvement indices between Covert and Overt treatments, across group configurations. Expert ratings of collaborative notes show that overtly signaling political affiliation significantly reduces the improvement in helpfulness for Democrat-Democrat teams, with this effect being significant at the 5% confidence level. No similar effect is observed for other group configurations.

Tweet	**Group**	**Category**	**Covert** $\mu(\sigma)$	**Overt** $\mu(\sigma)$	**T-value**	***p*-value (parametric)**	***p*-value[1] (non-parametric)**
Democrat	Homophilic (DD)	I_E	0.88 (1.5)	-0.167 (1.3)	2.2	0.0361*	0.0121*
		I_D	0.06 (1.0)	-0.112 (1.3)	0.4	0.6723	0.325
		I_R	1.03 (1.5)	-0.339 (1.4)	2.8	0.0079**	0.0017**
	Mixed (DR)	I_E	0.93 (2.1)	0.083 (1.3)	1.4	0.1713	0.0807
		I_D	0.53 (1.5)	-0.239 (1.1)	1.7	0.1045	0.0418*
		I_R	0.9 (1.4)	0.317 (1.1)	1.4	0.1798	0.0784
	Heterophilic (RR)	I_E	0.98 (1.6)	0.778 (1.6)	0.3	0.7355	0.3612
		I_D	0.33 (1.2)	0.126 (1.1)	0.5	0.6296	0.3073
		I_R	0.39 (1.8)	0.123 (1.2)	0.5	0.6299	0.3138
Republican	Homophilic (RR)	I_E	0.08 (1.2)	-0.028 (1.3)	0.2	0.806	0.4003
		I_D	0.33 (1.2)	-0.532 (1.0)	2.3	0.0275*	0.0064**
		I_R	0.36 (1.3)	-0.379 (1.1)	1.8	0.0788	0.0297*
	Mixed (DR)	I_E	0.98 (1.2)	0.426 (1.7)	1.1	0.2586	0.128
		I_D	1.09 (1.3)	0.546 (1.3)	1.2	0.2279	0.1052
		I_R	0.29 (1.4)	-0.022 (1.2)	0.7	0.4653	0.2209
	Heterophilic (DD)	I_E	0.75 (1.3)	-0.13 (1.5)	1.9	0.0673	0.0255*
		I_D	0.24 (1.0)	-0.062 (1.6)	0.7	0.5023	0.2488
		I_R	0.34 (0.9)	0.074 (1.3)	0.7	0.4772	0.2329

[1] Results are based on two-sample hypothesis tests using the Bootstrap method

(*) p<0.05; (**) p<0.01; (***) p<0.001

conducting an online experiment to evaluate the effectiveness of community-based approaches in countering misinformation and disinformation. The study examines how group composition and the visibility of party affiliation influence the accuracy of fact-checking.

Our work aimed to evaluate the effectiveness of community-based approaches in combating misinformation and disinformation on social media platforms, with a particular focus on political content. First, we find that teams, on average, produce more helpful notes than individuals, supporting our first hypothesis about the positive impact of collaboration on fact-checking.

Second, our findings underscore a nuanced relationship between group composition and political affiliation of content source. While we expected heterogeneous groups (DR) to outperform homogeneous groups (DD and RR, we found the opposite, we found that diversity is an asset when evaluating Republican tweets but not for Democrat tweets assessment, that is, homogeneous teams showed higher effectiveness when evaluating Democratic tweets, while unexpectedly, heterogeneous teams excelled in assessing Republican tweets.

Finally, the quality of collaborative notes declined when participants were explicitly aware of each other's political affiliations, reversing the collective intelligence effect. In contrast, when political affiliations were covert, teams produced higher-quality notes, demonstrating the benefits of anonymity in collaborative fact-checking.

We showed that teams generally outperform individuals in evaluating and annotating misleading political content. This finding reinforces established literature on the benefits of collaboration in complex problem-solving and suggests that community-based solutions could enhance the quality of information on social media platforms [26]. From the perspective of opinion diversity, heterogeneous groups were initially expected to outperform homogeneous groups. However, the results challenge this assumption: while teams with overt political affiliations did not perform as well, those with covert affiliations produced better outcomes. This result aligns with the notion that political identity can influence the objectivity and collaborative dynamics within groups, potentially leading to bias and reduced quality of fact-checking in mixed-affiliation teams. Additionally, it suggests that the anticipation of conflict may lead individuals to disengage from the task or conceal their political views, consistent with the Spiral of Silence theory [46, 62].

The insights from this study have practical implications for the design and implementation of community-based fact-checking systems on social media platforms. Our findings suggest that leveraging collective intelligence through collaborative efforts can enhance the quality of fact-checking. However, careful consideration of group composition and the visibility of political affiliations is crucial to maximize the effectiveness of these systems. Platforms may benefit from encouraging cross-partisan collaboration while maintaining anonymity regarding political affiliations to reduce bias and improve the impartiality of fact-checking outcomes. Future research should further explore strategies to optimize the composition and dynamics of collaborative fact-checking teams, ultimately contributing to more accurate and reliable information dissemination in the digital age.

Chapter 4

Exploring the impact of online collaborative environments on the willingness to express opinions: A Spiral of Silence perspective

4.1 Introduction

This research delves into the intricate relationship between human interaction and the processes of opinion formation and expression. While opinion dynamic models typically focus on interactions among individuals with differing viewpoints as drivers of opinion change, they often overlook how these interactions can influence behavior without necessarily altering core beliefs. Understanding how individuals respond to feedback and how their choices impact others' behavior is increasingly vital as more people globally engage with political content online. This study investigates how our inherent need for connection influences our decisions to share opinions on contentious topics in online discussions. To address this issue, we propose an online experiment to investigate whether collaborative online environments exacerbate individuals' fear of isolation, thereby lowering willingness to express their true opinions publicly. In our experiment, participants used an online interface to provide opinions on controversial topics, with the option to share or conceal their views with their connections. Participants were encouraged to form connections and faced penalties if others disconnected from them or rejected their attempts to connect. Our findings support our hypothesis that individuals conceal their opinions in response to negative feedback.

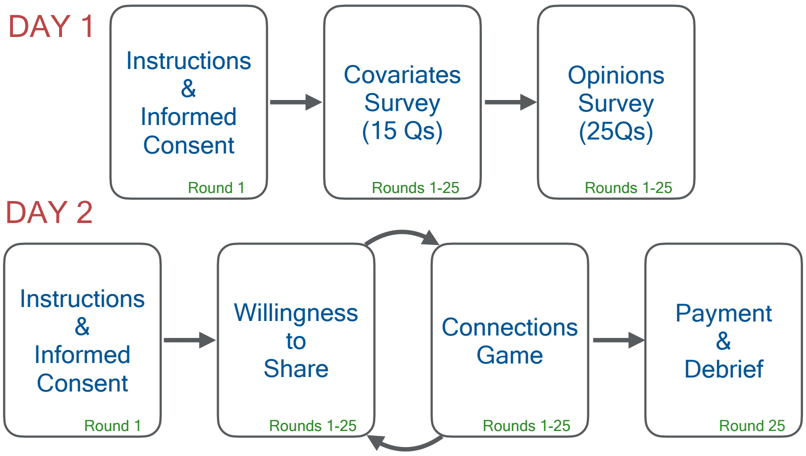


Figure 4.1: Experimental workflow

4.2 Research design

The aim of this work is to investigate whether online settings that incentivize connections with others heighten individuals' fear of isolation, thereby making them less likely to share their true opinions on controversial topics, particularly when their views differ from those of others. Specifically, we seek to experimentally test the main premise behind the Spiral of Silence [46].

To achieve this, we recognize that silence or self-censorship can manifest in two ways. First, by explicitly choosing not to share a view, or altering one's opinion to appear in agreement with others. To measure this concept, we designed a two-day experiment (see Figure 4.1 for a diagram of the experiment) where participants first complete an opinion questionnaire in a setting free from the influence of having to share their views publicly. Separating the measure of opinion from the interactive portion of the experiment increases the likelihood that participants answer genuinely. On the second day, participants were given the opportunity to revise their opinions before sharing them publicly. They were also asked whether they wanted to share their opinions with their connections, serving as an explicit measure of their willingness to share. We found that participants seldom chose to change their opinions (8% changed opinions), with opting not to share them being a more frequent choice (20% unshared opinions). We use willingness to share an opinion as our main outcome variable. See Table C.2 in Appendix for question wordings.

On the first day, participants completed a 16-question survey designed to assess six personality traits associated with a person's willingness to share opinions. The survey focused on fear of isolation—the primary predictor variable in this experiment—along with other factors influencing public opinion expression. These factors include predisposition to share opinions publicly [43], communication apprehension, defined as an individual's comfort with actual or anticipated oral communication [110], the psychological need to evaluate, which reflects a tendency to form strong opinions on various issues [143], as well as measures of political

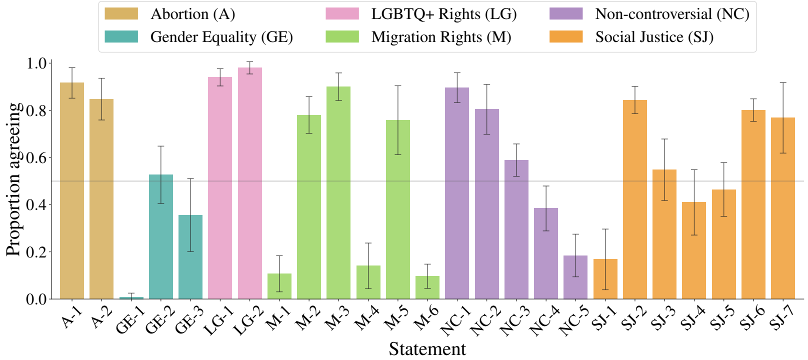


Figure 4.2: **Average agreement for each statement across six topics.** The height of each bar represents the mean proportion of participants agreeing with the statement, while the error bars indicate the standard error of the mean across sessions, reflecting variability in agreement levels.

engagement and online commenting behavior [76]. See Table C.1 in Appendix for variable definitions and question wordings.

Additionally, participants shared their opinions on 25 statements using a binary Agree/Disagree scale, addressing six topics: Abortion (A), Gender Equality (GE), LGBTQ+ Rights (LG), Migration Rights (M), Social Justice (SJ), and Non-controversial issues (NC). Refer to Table 4.1 for the wording of these statements. On Day 1, the statement order was randomized for each participant, and on Day 2, it was randomized per session. Figure 4.2 presents the average proportion of participants agreeing with each statement, as well as the variability in agreement levels (indicated by the error bars representing the standard error of the mean). Statements under LGBTQ+ Rights issues and question 1 on Gender Equality (GE-1) display small error bars, indicating consistent responses across sessions. In contrast, statements regarding a woman reducing paid work for her family (GE-2 and GE-3), the impact of immigrants on the Spanish economy (M-5), and government action to address income inequality (SJ-7) exhibit greater variability, suggesting more diverse and polarized opinions on these topics.

A second set of factors related to a person's willingness to share opinions was assessed alongside the opinion questionnaire. Participants answered questions on attitude certainty, which refers to the strength of conviction individuals have in their attitudes, or in other words, the extent to which an attitude is resistant to change [111, 144]; issue importance, both to oneself (day 1) and to one's social connections (day 2) [109]; and majority climate, referring to the individual perception of whether an opinion aligns with the majority view [111]. See Table C.3 in Appendix for variable definitions and question wordings.

Finally, we developed an interactive interface to allow participants to engage with others assigned to the same session. The purpose of this stage was to simulate a live exchange of opinions and provide clear signals of positive or negative feedback. On average, each session

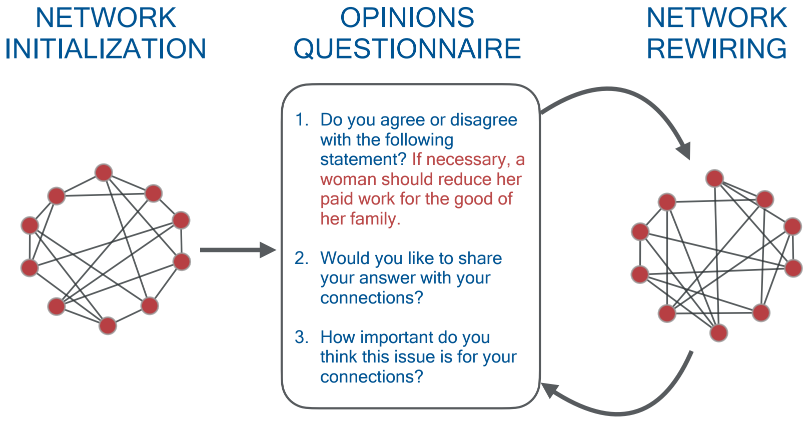

NETWORK INITIALIZATION
OPINIONS QUESTIONNAIRE
NETWORK REWIRING
1. Do you agree or disagree with the following statement? If necessary, a woman should reduce her paid work for the good of her family.
2. Would you like to share your answer with your connections?
3. How important do you think this issue is for your connections?

Figure 4.3: Illustration of the intended flow of the interactive session of the experiment.

included 25 active participants, and a total of 10 sessions were conducted. Initially, connections were randomized to form a random regular network with an average degree of 5. Interactions were then designed to allow for network rewiring, as illustrated in Figure 4.3. Additional statistics, such as the average number of connections per round, requests made, and disconnections, are provided in Figure C.4 in the Appendix.

The interactive stage aimed to let participants exchange opinions and learn about their connections' views on the chosen topics. An example of the interface is shown in Figure C.3 in the Appendix. Positive feedback was operationalized by allowing participants to send multiple connection requests, with successful connections earning both participants a payment of 100 tokens. Each connection request cost 5 tokens to send, and participants were given a budget of 20 tokens per round. Once a connection was established, participants could view each other's opinions for the current and all previous rounds during which the connection existed.

Negative feedback was operationalized through the ability to deny requests or disconnect from current connections. If a request was denied, the sender did not receive the token they paid to send it. When a connection was terminated, the disconnected participant was penalized with a deduction of 5 tokens from their total budget. Payoffs were calculated at the end of each round and accumulated over the session.

We recruited 400 participants for the experiment (40 participants per session across 10 sessions), most of whom completed the day 1 surveys. Of these, 298 returned on day 2, indicating a 25.5% dropout rate from day 1 to day 2. Among the 298 participants, 252 completed all rounds on day 2, resulting in a 15.4% dropout rate on day 2 and a 37% overall dropout rate. Refer to Table C.4 in the Appendix for a session-by-session breakdown of these numbers.

Table 4.1: **Opinion statement wordings.** Participants were asked to indicate their opinions on 25 statements across six topics using a binary Agree/Disagree scale. The topics included Abortion (A), Gender Equality (GE), LGBTQ+ Rights (LG), Migration Rights (M), Social Justice issues (SJ), and Non-controversial issues (NC).

Code	**Statement**
A-1	By law, abortion must be allowed under certain circumstances
A-2	Abortion is a human right
GE-1	The role of women in society should be limited to household chores and care of others
GE-2	A woman should not reduce her paid work for the sake of her family, even if necessary
GE-3	If necessary, a woman should reduce her paid work for the good of her family
LG-1	I would not feel ashamed if a close relative was homosexual
LG-2	Homosexual couples should have the same rights as heterosexual couples
M-1	Immigrants come to Spain only to benefit from social benefits and services
M-2	Immigration enriches Spanish culture
M-3	Legal immigrants should have the same rights as Spanish citizens
M-4	Immigrants endanger Spanish culture
M-5	Immigrants coming to Spain contribute positively to the Spanish economy
M-6	Legal immigrants must not have the same rights as Spanish citizens
NC-1	The teaching of a language other than Spanish should be mandatory in basic education
NC-2	A true Spanish omelette (tortilla de patatas) has onions
NC-3	Zoos contribute to the protection of endangered species
NC-4	You can't call something that has pineapple on it a pizza
NC-5	Basketball is a more complete sport than swimming
SJ-1	Wealthy families should have access to higher level of services
SJ-2	A just society takes care of the poor and needy, regardless of what they contribute to society
SJ-3	Large differences in people's incomes are acceptable in order to adequately reward differences in talents and efforts
SJ-4	Large differences in income are not acceptable, even if they reflect differences in talent and effort
SJ-5	The Spanish government spends too little on social benefits and services
SJ-6	The government should do everything possible to improve the social and economic position of minority groups
SJ-7	The government must take measures to reduce the differences in income levels

4.3 Results and discussion

Our data follows a hierarchical structure given our two sets of predictor variables—personality characteristics and round-specific covariates. The lowest level (level 1) consists of round-specific variables for each participant, while the highest level (level 2) reflects participant-level characteristics measured through the personality traits survey. To make use of this structure, we specified a logistic multilevel regression model to predict the odds of the willingness to share (Y). This method addresses the lack of independence among observations by incorporating random intercepts, and capturing variability between participants [145].

Table 4.2 summarizes the results for three model specifications: Model 1 includes participant-level variables only, Model 2 focuses solely on round-level variables, and Model 3 incorporates all variables. The results indicate that, in both Models 1 and 3, the predisposition to share is the only participant-level variable that significantly affects the odds of sharing. As expected, an increase in predisposition to share is positively associated with an increased likelihood of sharing.

Including this variable in the model serves two purposes. First, it accounts for individual differences in predisposition to share, helping to isolate the specific relationships between other independent variables of interest, such as fear of isolation and attitude certainty. Second, it demonstrates a positive relationship between survey-based measures of willingness to share opinions and actual sharing behavior—at least in online settings. Like previous work [43], this finding supports the reliability of hypothetical survey questions in assessing willingness to share opinions and emphasizes the need to account for individual predispositions in analyses of sharing behavior.

The Spiral of Silence theory suggests that individuals evaluate the opinion climate—how closely their views align with perceived public opinion—when deciding whether to express their opinions [110]. In this study, we operationalize this theory by including measures for the perceived opinion climate (Majority Climate), the actual opinion climate (measured by the number of neighbors in disagreement, N Neighbors Disagree), and negative feedback (quantified as the number of disconnections received in a given round, N Disconnects).

Our findings indicate that only the perception of the majority climate significantly influences the decision to share opinions. Specifically, an increase in the belief that one's opinion aligns with the majority increases the odds of sharing. While this positive relationship aligns with the fundamental premise of the Spiral of Silence, the novel insight from our study is that perceived opinion climate has a stronger influence on sharing decisions than the actual opinion climate. This highlights the primacy of perception over reality in shaping public expression of opinions. These results contrast previous modeling efforts that underscore the importance of network effects for a Spiral of Silence effect to emerge [104]. Nevertheless, it is possible that

Table 4.2: Willingness to share, (Y)

	Dependent variable:		
	Willingness to Share, (Y)		
	(1)	(2)	(3)
Participant Level			
Predisposition to Share	0.223^{**}		0.232^{**}
	(0.081)		(0.084)
Fear of Isolation	-0.035		-0.032
	(0.097)		(0.101)
Communication Apprehension	0.006		0.015
	(0.124)		(0.129)
Need to Evaluate	0.170		0.085
	(0.131)		(0.137)
Engaged in Politics	0.112		0.104
	(0.073)		(0.076)
Active Social Media	0.167		0.181
	(0.182)		(0.190)
Round Level			
Majority Climate		0.165^{***}	0.166^{***}
		(0.043)	(0.043)
Attitude Certainty		0.296^{***}	0.289^{***}
		(0.044)	(0.044)
Issue Importance for Connections		0.266^{***}	0.265^{***}
		(0.037)	(0.037)
N Neighbors Disagree		0.003	0.003
		(0.006)	(0.006)
N Disconnects		0.005	0.006
		(0.032)	(0.032)
LGBTQ+ Rights		0.535^{*}	0.545^{*}
		(0.225)	(0.225)
Abortion Rights		-0.762^{***}	-0.753^{***}
		(0.167)	(0.167)
Gender Equality		-0.978^{***}	-0.973^{***}
		(0.142)	(0.142)
Migration		-0.636^{***}	-0.630^{***}
		(0.122)	(0.122)
Social Justice		-0.985^{***}	-0.980^{***}
		(0.119)	(0.119)
N Round		-0.024^{***}	-0.024^{***}
		(0.005)	(0.005)
Constant	-0.104	-0.132	-1.698^{*}
	(0.761)	(0.247)	(0.825)
Observations	5,992	5,992	5,992
Log Likelihood	-2,731.183	-2,583.446	-2,577.699
Akaike Inf. Crit.	5,478.366	5,192.892	5,193.397
Bayesian Inf. Crit.	5,531.952	5,279.968	5,320.663
Note:			*p<0.05; **p<0.01; ***p<0.001

our experimental setting—particularly the time constraints—did not adequately allow participants to experience the full impact of network effects. Future research aiming to examine the influence of network effects on the Spiral of Silence would benefit from an experimental design that places greater emphasis on capturing these dynamics.

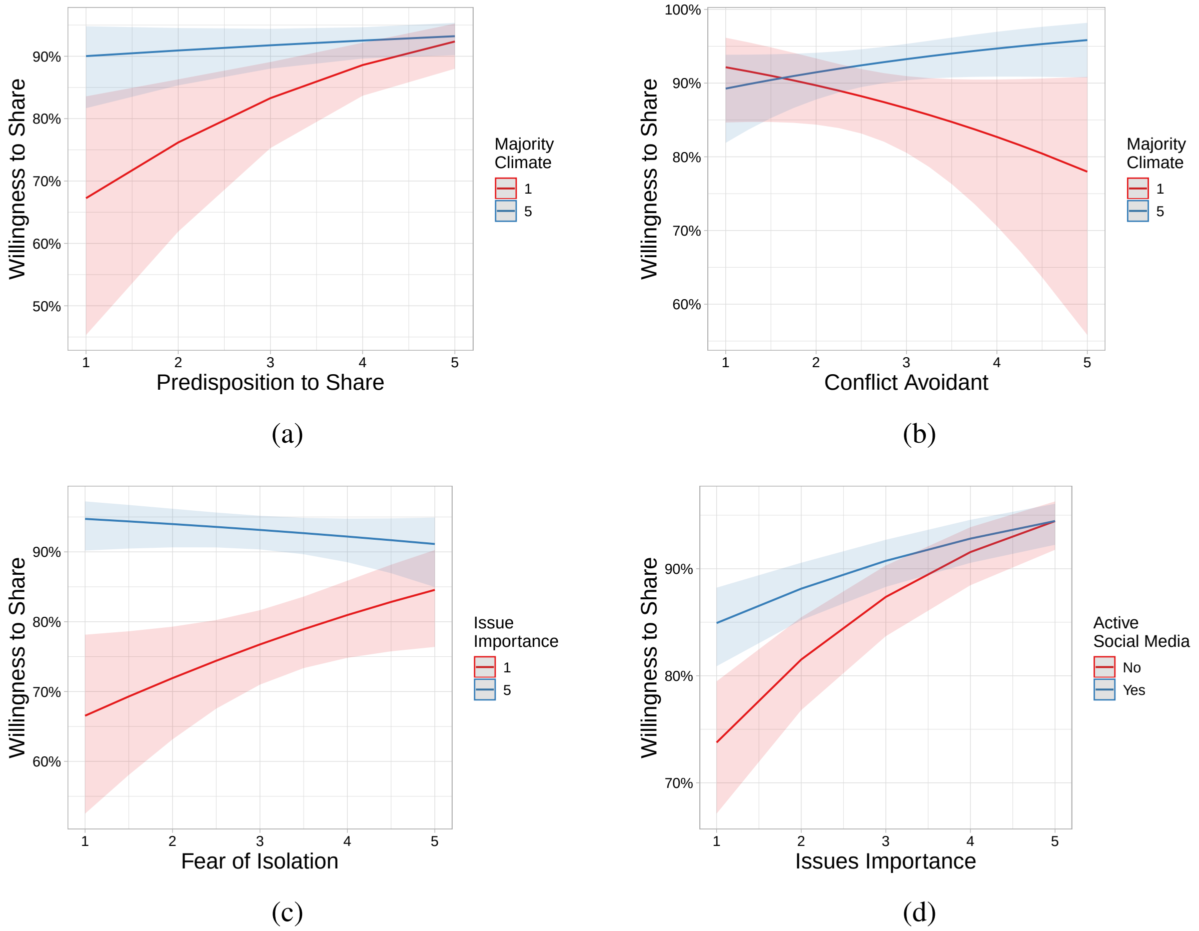
Both theoretical formulations and empirical studies have shown that the Spiral of Silence applies primarily to individuals with low to moderate attitude certainty [111], as those with high attitude certainty are less fearful of isolation and, consequently, more likely to share their opinions publicly. Our findings confirm this insight, showing that high attitude certainty is associated with increased odds of publicly sharing one's opinion.

Issue importance is strongly correlated with attitude certainty but represents a distinct concept. While attitude certainty pertains to how confident an individual is in their own position, issue importance reflects the relative significance one assigns to a specific topic [109]. Individuals who attribute greater importance to a topic are often more likely to exhibit high attitude certainty, though this is not always the case. Regarding willingness to speak, higher issue importance increases the likelihood of sharing opinions.

Our analysis found that issue importance to the individual (ego) was strongly correlated with attitude certainty, whereas issue importance for connections was not correlated with any independent variable (see Figure C.2 in the Appendix). As a result, we excluded issue importance to ego from our model specification. We found that an increase in the perceived importance of the issue to connections was associated with higher odds of sharing one's opinions with others, confirming theoretical predictions.

Based on the Spiral of Silence theory, we hypothesized that fear of isolation would be negatively associated with the willingness to share opinions [44, 108, 110, 111]. Similarly, individuals with high communication apprehension—those who feel uncomfortable engaging in conversations—were expected to be less likely to participate in discussions on controversial topics and, consequently, less likely to share their opinions [110]. In contrast, we anticipated a positive relationship between the willingness to share and factors such as the need to evaluate, political engagement, and activity on social media [76].

However, none of these participant-level variables demonstrated significant associations with the odds of sharing. Notably, some variables became significant when interacting with round-specific factors, suggesting context-dependent influences on the willingness to share opinions, such as the interaction between communication apprehension and majority climate. Figure 4.4, panel 4.4b (see Table 4.3 for complete results), illustrates that the predicted probability of sharing one's opinion decreases as communication apprehension increases, especially when the opinion is perceived to hold minority status. Conversely, when an opinion is perceived to hold majority status, the predicted probability of sharing one's opinion increases with higher communication apprehension, though this increase is more modest compared to the previously



(a)

(b)

(c)

(d)

Figure 4.4: Significant cross-level interactions.

described decrease.

This finding suggests that communication apprehension shapes how the perception of opinion climate affects the willingness to share, partially confirming theoretical expectations. Specifically, individuals who feel uncomfortable engaging in conversations and perceive their opinion to be in the minority are less likely to share their opinions publicly.

The interaction between predisposition to share and majority climate is significant, as shown in Table 4.3. Panel 4.4a shows that while willingness to share generally rises with an increased predisposition to share, the increase is notably steeper when an opinion is perceived to hold minority status. This suggests that individuals with a high predisposition to share are less influenced by perceptions of the opinion climate.

The interaction between fear of isolation and issue importance to connections was also found to be significant. Panel 4.4c demonstrates that when an issue is perceived as having little importance to one's connections, an increase in fear of isolation corresponds to a higher predicted probability of sharing opinions publicly. Conversely, when the issue is considered highly important to connections, greater fear of isolation results in a decreased willingness to share. This

Table 4.3: Willingness to share answer, Y

	Dependent variable:			
	Share_Y2			
	(1)	(2)	(3)	(4)
Participant Level				
Predisposition to Share	0.528**	0.235**	0.220**	0.235**
	(0.174)	(0.078)	(0.085)	(0.078)
Fear of Isolation	−0.039	0.352*	−0.037	0.079
	(0.101)	(0.150)	(0.101)	(0.096)
Conflict Avoidant	0.052	−0.026	−0.438	−0.018
	(0.132)	(0.120)	(0.251)	(0.120)
Need to Evaluate	0.083	0.143	0.083	0.143
	(0.138)	(0.126)	(0.138)	(0.126)
Engaged in Politics	0.110	0.097	0.111	0.099
	(0.077)	(0.070)	(0.077)	(0.070)
Active Social Media	0.188	0.398*	0.175	0.868***
	(0.190)	(0.181)	(0.190)	(0.260)
Round Level				
Majority Climate	0.455**	0.174***	−0.225	0.172***
	(0.166)	(0.043)	(0.169)	(0.043)
Attitude Certainty	0.297***	0.278***	0.295***	0.280***
	(0.046)	(0.045)	(0.046)	(0.045)
Issue Importance for Connections	0.267***	0.648***	0.267***	0.450***
	(0.037)	(0.133)	(0.037)	(0.059)
N Neighbors Disagree	0.003	0.004	0.003	0.003
	(0.006)	(0.006)	(0.006)	(0.006)
N Disconnects	0.009	0.008	0.009	0.006
	(0.032)	(0.032)	(0.032)	(0.032)
LGBTQ+ Rights	0.543*	0.585**	0.537*	0.578*
	(0.226)	(0.225)	(0.226)	(0.225)
Abortion Rights	−0.755***	−0.733***	−0.765***	−0.731***
	(0.168)	(0.167)	(0.168)	(0.167)
Gender Equality	−0.983***	−0.979***	−0.987***	−0.962***
	(0.143)	(0.141)	(0.143)	(0.141)
Migration	−0.654***	−0.614***	−0.647***	−0.610***
	(0.124)	(0.121)	(0.124)	(0.121)
Social Justice	−0.990***	−0.979***	−0.988***	−0.975***
	(0.120)	(0.118)	(0.120)	(0.118)
Round	−0.025***	−0.024***	−0.025***	−0.024***
	(0.005)	(0.005)	(0.005)	(0.005)
Cross-Level Interactions				
Predisposition to Share ×	−0.085*			
Majority Climate	(0.042)			
Fear of Isolation ×		−0.098*		
Issue Important for Connections		(0.041)		
Conflict Avoidant ×			0.139*	
Majority Climate			(0.062)	
Active Social Media ×				−0.173*
Majority Climate				(0.072)
Constant	−2.830**	−3.346***	−0.384	−2.817***
	(1.021)	(0.833)	(1.016)	(0.769)
Observations	5,992	5,992	5,992	5,992
Log Likelihood	-2,572.670	-2,558.663	-2,572.239	-2,558.603
Akaike Inf. Crit.	5,189.339	5,161.326	5,188.478	5,161.207
Bayesian Inf. Crit.	5,336.699	5,308.686	5,335.838	5,308.567
Note:			*p<0.05; **p<0.01; ***p<0.001	

finding suggests that the theoretical relationship between fear of isolation and willingness to share becomes relevant only when the issue holds significance for others. Conversely, when the issue is perceived as unimportant to connections, fear of isolation appears to motivate greater sharing.

Finally, panel 4.4d shows that social media activity influences willingness to share opinions when an issue is unimportant to connections. However, as the perceived importance of the issue increases, the effect of social media activity dissipates.

4.4 Conclusions

Our results reveal positive associations between high predisposition to share, perception of majority climate, high attitude certainty, and the perception of high issue importance for connections with the likelihood of sharing an opinion publicly. While we did not find any effect of fear of isolation on the odds of sharing, the observed associations align with core premises of the Spiral of Silence theory. Specifically, individuals take the prevailing opinion climate into account when deciding whether to voice their opinions [110], those with high attitude certainty are more likely to express their opinions publicly [111], and greater issue importance increases the likelihood of sharing opinions [109].

We found that while most participant-level variables are not significant on their own, they become significant when in interaction with round-level variables. Specifically, individuals with a strong predisposition to share are less influenced by their perceptions of the opinion climate. Communication apprehension shapes how perceptions of the opinion climate affect willingness to share, with individuals who are uncomfortable in conversations and perceive their opinion as a minority being less likely to express it publicly. The relationship between fear of isolation and willingness to share is significant only when the issue is important to others; when the issue is deemed unimportant to connections, fear of isolation appears to encourage greater sharing. Finally, social media activity influences the willingness to share opinions when an issue is unimportant to connections, but this effect weakens as the perceived importance of the issue increases.

It is well-established that individuals' online activity is influenced not only by their personal convictions but also by the opinions expressed within their social networks [62]. However, the reliance on perception rather than reality–as our results suggest–when deciding to publicly express opinions complicates efforts to understand the true drivers of online political expression. This issue extends beyond recognizing that social media fails to provide an accurate and unbiased reflection of public opinions and beliefs, rendering it unsuitable as a substitute for survey data [62, 76]. It also underscores the risks associated with exploiting individual vulnerabilities on social media. For instance, just as misinformation has been shown to exploit moral outrage—a combination of disgust and anger—to spread online [146], perceptions of prevailing opinions can also be manipulated. Certain ideas may be artificially amplified to create a false sense of prominence. This distortion can mislead individuals about the majority climate, potentially resulting in self-censorship and reducing the diversity of voices in public discourse. It is imperative for researchers to investigate whether this phenomenon occurs and, if so, to understand its implications for democracy.

Chapter 5

Conclusions

Understanding how individuals respond to social influence is crucial for studying collective political behavior in online spaces. While much of the existing research on public forum opinions focuses on social feedback [106], it often neglects the role of human interactions in fostering self-censorship [62, 104]. This work investigates political deliberation in online environments, testing the hypothesis that individuals may withhold minority opinions in public due to fears of encountering toxic behavior. To explore how digital media encourages or hinders diverse political expression, this work combines complementary research methods: data analysis of social media (Chapter 2) and online social experiments (Chapters 3 and 4). These approaches provide unique insights while addressing each other's limitations.

The approaches used in this work are complementary, leveraging the strengths of both observational data and experimental methods. Social media platforms like YouTube provide valuable data for identifying large-scale patterns and trends in collective behavior, though the influence of algorithms, user activity, and moderation practices complicates causal inference and data completeness [4, 62]. Synchronous online social experiments address these limitations by offering a controlled environment to test hypotheses and study behaviors under specific conditions. By bridging micro-level individual actions and macro-level patterns, these experiments capture the emergent dynamics of interactions in real-time, complementing the broader insights from social media data [4, 67, 68].

In Chapter 2, we analyze YouTube comments from the 2020 US presidential elections to investigate the prevalence of toxic and insulting language in online political discussions and its impact on conversation dynamics and disengagement. We found that toxic and insulting content is especially common during periods of political tension. Toxic top-level comments are more likely to provoke similarly negative replies, creating a cyclical pattern of toxicity and escalating hostility within discussions.

Additionally, employing hidden Markov models, we identify a latent state linked to toxicity-driven disengagement, marked by reduced user activity and an increased likelihood of posting toxic content. This dynamic fosters environments dominated by extreme and antisocial behavior. These patterns, confirmed across various video datasets and for insulting language, underscore the significant role of toxic interactions in shaping political discourse. Specifically, they discourage participation from marginalized and moderate voices, emphasizing the need to address self-censorship dynamics in online political environments.

This study has several limitations related to data completeness, representativeness, and methodological differences. First, we assume that all conversations end due to toxic or insulting behavior, which may lead to an overestimation of the effect of negative sentiment on disengagement. Second, the channels analyzed diligently moderate comments, likely excluding the most problematic ones, thus, limiting the generalizability of our findings. Additionally, while our study expands on prior research by analyzing both toxic and insulting behaviors, direct comparisons to works such as Avalle et al. [80] are challenging due to differing definitions of a conversation. Our approach focuses on coherent, time-bound exchanges centered on a single top-level comment and its replies, ensuring topic-specific interactions, whereas Avalle et al. [80] consider an entire video thread, assuming chronological engagement that may not align with user behavior on YouTube. These definitional and methodological differences, coupled with the potential influence of video content over sequential comment interactions, complicate comparative analyses and highlight the need for caution in interpreting results.

In Chapter 3, we conducted an experiment in which participants collaboratively evaluated political content under conditions of either overt or covert political affiliations. The findings show that collaboration enhances fact-checking quality, but overt political affiliations diminish the effectiveness of outputs, highlighting the value of anonymity in collaborative environments. Additionally, group diversity proved beneficial for evaluating Republican tweets but not for assessing Democrat tweets. Overall, the results suggest that anticipating conflict may prompt individuals to withdraw from the task or hide their political views, aligning with the Spiral of Silence theory.

Although Chapter 3 does not directly analyze the dynamics of political expression in online forums—the central focus of this dissertation—it offers two key contributions. First, the experiment creates a scenario where political expression is necessary to complete a specific task, prompting participants to consider their political views when collaborating with others. Contrary to expectations that diverse groups would outperform homogeneous ones, the results reveal that teams with covert political affiliations achieved better outcomes than those with overt affiliations, suggesting that conflict anticipation may cause disengagement or self-censorship, aligning with the Spiral of Silence theory. Second, the study provided a platform to refine interactive experiment design, including coding processes and participant recruitment strategies, offering valuable insights for the subsequent chapter's experiment.

In Chapter 4, we design an experiment using an online interface to test the Spiral of Silence theory. While we found no significant relationship between fear of isolation and willingness to share opinions, other findings support core aspects of the theory. Specifically, individuals with high communication apprehension are less likely to participate in controversial discussions, whereas high attitude certainty and issue importance increase the likelihood of opinion sharing.

All in all, we have confirmed that individuals' online activity is shaped by both personal beliefs and the opinions expressed within their social networks [62]. However, our findings suggest that decisions to publicly express opinions are influenced more by perception than reality, complicating efforts to understand the true drivers of political expression. This reliance on perception underscores that social media fails to provide an accurate reflection of public opinion, rendering it unsuitable as a substitute for survey data [62, 76]. Furthermore, it highlights the risks of exploiting vulnerabilities on social media. Just as misinformation has been found to exploit moral outrage to spread online [146], perceptions of dominant opinions can also be manipulated. Certain ideas may be artificially amplified to give a false impression of prominence. This distortion can mislead individuals about the prevailing majority opinion, potentially leading to self-censorship and diminishing the diversity of voices in public discourse. It is crucial for researchers to examine whether this phenomenon occurs and to understand its potential consequences for democracy.

The Spiral of Silence is widely recognized as an incomplete explanation of the drivers behind political expression [47,62,147]. However, it offers valuable insights into the interplay between social influence and political expression and served as a guiding framework to shape the scope and direction of this work. The primary limitation of the Spiral of Silence theory is its failure to account for social processes that exhibit what Galesic et al. [67] describe as ”collective adaptation”. This concept refers to the dynamic process by which groups adjust their behaviors, strategies, and structures in response to evolving challenges. Unlike static tasks, social groups face changing problem landscapes that require flexibility and the capacity to address multiple issues simultaneously [67, 68]. Within this framework, paradoxes such as a minority opinion holding majority status, as posited by the Spiral of Silence theory, can be better understood. This is because collectives often encounter interference from existing integration strategies and social networks, which may be outdated or misaligned with current issues [67].

As the digital transformation of society has led to the emergence of new networked structures and collective behaviors, interdisciplinary collaboration is needed for effective understanding of the underlying mechanisms [68]. To effectively address complex societal challenges such as radicalization, polarization, misinformation, and group violence, future research should broaden its traditional individual-centric focus to encompass a more comprehensive understanding of human collectives [68]. This shift is essential, as human beliefs, emotions, decisions, and behaviors are deeply influenced by collective dynamics [67].

Bibliography

[1] S. González-Bailón, D. Lazer, P. Barberá, M. Zhang, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Freelon, M. Gentzkow, A. Guess, *et al.*, "Asymmetric ideological segregation in exposure to political news on facebook," *Science*, vol. 381, no. 6656, pp. 392–398, 2023.

[2] A. Guess, N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, *et al.*, "How do social media feed algorithms affect attitudes and behavior in an election campaign?," *Science*, vol. 381, no. 6656, pp. 398–404, 2023.

[3] P. Boczkowski, "The mutual shaping of technology and society in videotex newspapers: Beyond the diffusion and social shaping perspectives," *The information society*, vol. 20, no. 4, pp. 255–267, 2004.

[4] J. Bak-Coleman, M. Alfano, W. Barfuss, C. Bergstrom, M. Centeno, I. Couzin, J. Donges, M. Galesic, A. Gersick, J. Jacquet, *et al.*, "Stewardship of global collective behavior," *Proceedings of the National Academy of Sciences*, vol. 118, no. 27, p. e2025764118, 2021.

[5] J. Henrich, *The Secret of Our Success*. Princeton: Princeton University Press, 2017.

[6] S. Galam, "Contrarian deterministic effects on opinion dynamics:"the hung elections scenario"," *Physica A: Statistical Mechanics and its Applications*, vol. 333, pp. 453–460, 2004.

[7] A. B. Kao and I. D. Couzin, "Decision accuracy in complex environments is often maximized by small group sizes," *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1784, p. 20133305, 2014.

[8] M. Galesic, D. Barkoczi, and K. Katsikopoulos, "Smaller crowds outperform larger crowds and individuals in realistic task conditions.," *Decision*, vol. 5, no. 1, p. 1, 2018.

[9] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in *European conference on principles of data mining and knowledge discovery*, pp. 259–271, Springer, 2006.

[10] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st international conference on World Wide Web*, pp. 519–528, 2012.

[11] K. Koltai, "Vaccine information seeking and sharing: How private facebook groups contributed to the anti-vaccine movement online," *AoIR Selected Papers of Internet Research*, 2020.

[12] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand, "Shifting attention to accuracy can reduce misinformation online," *Nature*, vol. 592, no. 7855, pp. 590–595, 2021.

[13] L. Chittka, P. Skorupski, and N. E. Raine, "Speed–accuracy tradeoffs in animal decision making," *Trends in ecology & evolution*, vol. 24, no. 7, pp. 400–407, 2009.

[14] G. Pennycook and D. G. Rand, "Examining false beliefs about voter fraud in the wake of the 2020 presidential election," *The Harvard Kennedy School Misinformation Review*, 2021.

[15] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[16] I. Rahwan, "Society-in-the-loop: programming the algorithmic social contract," *Ethics and information technology*, vol. 20, no. 1, pp. 5–14, 2018.

[17] A. Gimmler, "Deliberative democracy, the public sphere and the internet," *Philosophy & Social Criticism*, vol. 27, no. 4, pp. 21–39, 2001.

[18] B. Bressers and J. Hume, "Message boards, public discourse, and historical meaning: An online community reacts to September 11," *American Journalism*, vol. 29, no. 4, pp. 9–33, 2012.

[19] P. Lorenz-Spreen, L. Oswald, S. Lewandowsky, and R. Hertwig, "A systematic review of worldwide causal and correlational evidence on digital media and democracy," *Nature human behaviour*, vol. 7, no. 1, pp. 74–101, 2023.

[20] G. Wolfsfeld, E. Segev, and T. Sheafer, "Social media and the arab spring: Politics comes first," *The International Journal of Press/Politics*, vol. 18, no. 2, pp. 115–137, 2013.

[21] A. Ansari, "The role of social media in iran's green movement (2009-2012)," *Global Media Journal-Australian Edition*, vol. 6, no. 2, pp. 1–6, 2012.

[22] L. Manikonda, G. Beigi, H. Liu, and S. Kambhampati, “Twitter for sparking a movement, reddit for sharing the moment:# metoo through the lens of social media,” *arXiv preprint arXiv:1803.08022*, 2018.

[23] M. Mundt, K. Ross, and C. M. Burnett, “Scaling social movements through social media: The case of black lives matter,” *Social media+ society*, vol. 4, no. 4, p. 2056305118807911, 2018.

[24] L. Asmelash, “How Black Lives Matter went from a hashtag to a global rallying cry,” *CNN*, 26 July 2020. https://edition.cnn.com/2020/07/26/us/black-lives-matter-explainer-trnd/index.html.

[25] S. Frenkel, “The storming of Capitol Hill was organized on social media,” *New York Times*, 1 January 2021. https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html.

[26] T. Yasseri and F. Menczer, “Can crowdsourcing rescue the social marketplace of ideas?,” *Communications of the ACM*, vol. 66, no. 9, pp. 42–45, 2023.

[27] Z. Papacharissi, “The virtual sphere: The internet as a public sphere,” *New media & society*, vol. 4, no. 1, pp. 9–27, 2002.

[28] B. Stark, D. Stegmann, M. Magin, and P. Jürgens, “Are algorithms a threat to democracy? The rise of intermediaries: A challenge for public discourse,” *Algorithm Watch*, 26 May 2020. https://algorithmwatch.org/de/wp-content/uploads/2020/05/Governing-Platforms-communications-study-Stark-May-2020-AlgorithmWatch.pdf.

[29] J. Stoyanovich, J. J. Van Bavel, and T. V. West, “The imperative of interpretable machines,” *Nature Machine Intelligence*, vol. 2, no. 4, pp. 197–199, 2020.

[30] B. Bimber, A. Flanagin, and C. Stohl, *Collective action in organizations: Interaction and engagement in an era of technological change*. Cambridge University Press, 2012.

[31] H. Margetts, P. John, S. Hale, and T. Yasseri, *Political turbulence: How social media shape collective action*. Princeton: Princeton University Press, 2016.

[32] G. Porumbescu, “Not all bad news after all? Exploring the relationship between citizens’ use of online mass media for government information and trust in government,” *International Public Management Journal*, vol. 20, no. 3, pp. 409–441, 2017.

[33] A. Guess, P. Barberá, S. Munzert, and J. Yang, “The consequences of online partisan media,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 14, p. e2013464118, 2021.

[34] S. Schumann, F. Thomas, F. Ehrke, T. Bertlich, and J. Dupont, “Maintenance or change? Examining the reinforcing spiral between social media news use and populist attitudes,” *Information, Communication & Society*, vol. 25, no. 13, pp. 1934–1951, 2022.

[35] S. Castaõ-Pulgarín, N. Suárez-Betancur, L. Tilano-Vega, and H. Herrera-Lṕez, “Internet, social media and online hate speech. Systematic review,” *Aggression and Violent Behavior*, vol. 58, p. 101608, 2021.

[36] B. Vidgen and T. Yasseri, “Detecting weak and strong islamophobic hate speech on social media,” *Journal of Information Technology & Politics*, vol. 17, no. 1, pp. 66–78, 2020.

[37] C. Bail, L. Argyle, T. Brown, J. Bumpus, H. Chen, F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, “Exposure to opposing views on social media can increase political polarization,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. 9216–9221, 2018.

[38] M. Yarchi, C. Baden, and N. Kligler-Vilenchik, “Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media,” *Political Communication*, vol. 38, no. 1-2, pp. 98–139, 2021.

[39] R. Bond, C. Fariss, J. Jones, A. Kramer, C. Marlow, J. Settle, and J. Fowler, “A 61-million-person experiment in social influence and political mobilization,” *Nature*, vol. 489, no. 7415, pp. 295–298, 2012.

[40] S. Boulianne, “Twenty years of digital media effects on civic and political participation,” *Communication Research*, vol. 47, no. 7, pp. 947–966, 2020.

[41] C. Beaudoin, “The internet’s impact on international knowledge,” *New Media & Society*, vol. 10, no. 3, pp. 455–474, 2008.

[42] C. Park and B. Kaye, “News engagement on social media and democratic citizenship: Direct and moderating roles of curatorial news use in political involvement,” *Journalism & Mass Communication Quarterly*, vol. 95, no. 4, pp. 1103–1127, 2018.

[43] D. A. Scheufele, J. Shanahan, and E. Lee, “Real talk: Manipulating the dependent variable in spiral of silence research,” *Communication research*, vol. 28, no. 3, pp. 304–324, 2001.

[44] J. Matthes, J. Knoll, and C. von Sikorski, “The “spiral of silence” revisited: A meta-analysis on the relationship between perceptions of opinion support and political opinion expression,” *Communication Research*, vol. 45, no. 1, pp. 3–33, 2018.

[45] M. Skoric, Q. Zhu, and N. Pang, "Social media, political expression, and participation in confucian asia," *Chinese Journal of Communication*, vol. 9, no. 4, pp. 331–347, 2016.

[46] E. Noelle-Neumann, "The spiral of silence a theory of public opinion," *Journal of communication*, vol. 24, no. 2, pp. 43–51, 1974.

[47] D. A. Scheufle and P. Moy, "Twenty-five years of the spiral of silence: A conceptual review and empirical outlook," *International journal of public opinion research*, vol. 12, no. 1, pp. 3–28, 2000.

[48] C. J. Glynn, "Public opinion as a normative opinion process," *Annals of the International Communication Association*, vol. 20, no. 1, pp. 157–183, 1997.

[49] T. Kuran, *Private truths, public lies: The social consequences of preference falsification*. Cambridge, MA: Harvard University Press, 1995.

[50] B. Latane and J. M. Darley, "Group inhibition of bystander intervention in emergencies," *Journal of personality and social psychology*, vol. 10, no. 3, p. 215, 1968.

[51] D. T. Miller and D. A. Prentice, "Collective errors and errors about the collective," *Personality and Social Psychology Bulletin*, vol. 20, no. 5, pp. 541–550, 1994.

[52] R. H. Sargent and L. S. Newman, "Pluralistic ignorance research in psychology: A scoping review of topic and method variation and directions for future research," *Review of General Psychology*, vol. 25, no. 2, pp. 163–184, 2021.

[53] R. E. Nisbett and Z. Kunda, "Perception of social distributions," *Journal of Personality and Social Psychology*, vol. 48, no. 2, p. 297, 1985.

[54] M. R. Leary, *Self-presentation: Impression management and interpersonal behavior*. United Kingdom: Routledge, 2019.

[55] S. Gavrilets, "Coevolution of actions, personal norms and beliefs about others in social dilemmas," *Evolutionary Human Sciences*, vol. 3, p. e44, 2021.

[56] K. Hampton, l. Rainie, W. Lu, M. Dwyer, I. Shin, and K. Purcell, "Social media and the spiral of silence," *Pew Research Center, Washington, D.C.*, 25 August 2014. https://www.pewresearch.org/internet/2014/08/26/social-media-and-the-spiral-of-silence/.

[57] C. P. Hoffmann and C. Lutz, "Spiral of silence 2.0: Political self-censorship among young facebook users," in *Proceedings of the 8th international conference on social media & society*, (New York, NY, USA), pp. 1–12, Association for Computing Machinery, 2017.

[58] G. Neubaum and N. C. Krämer, “Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media,” *Media psychology*, vol. 20, no. 3, pp. 502–531, 2017.

[59] E. A. Bäck, H. Bäck, A. Fredén, and N. Gustafsson, “A social safety net? rejection sensitivity and political opinion sharing among young people in social media,” *New Media & Society*, vol. 21, no. 2, pp. 298–316, 2019.

[60] H.-T. Chen, “Spiral of silence on social media and the moderating role of disagreement and publicness in the network: Analyzing expressive and withdrawal behaviors,” *New Media & Society*, vol. 20, no. 10, pp. 3917–3936, 2018.

[61] J. E. Settle, *Frenemies: How social media polarizes America*. Cambridge, UK: Cambridge University Press, 2018.

[62] W. S. Schulz, A. M. Guess, P. Barberá, S. Munzert, A. Gottlieb, A. Hughes, E. Remy, S. Shah, and A. Smith, “(Mis)representing ideology on twitter: How social influence shapes online political expression,” 2020. https://simonmunzert.github.io/meof/material/schulz-et-al-ideology-twitter-apsa.pdf.

[63] A. E. Marwick and D. Boyd, “I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience,” *New media & society*, vol. 13, no. 1, pp. 114–133, 2011.

[64] E. Katz, P. F. Lazarsfeld, and E. Roper, *Personal influence: The part played by people in the flow of mass communications*. New York: Routledge, 2017.

[65] N. Beauchamp, “Predicting and interpolating state-level polls using twitter textual data,” *American Journal of Political Science*, vol. 61, no. 2, pp. 490–503, 2017.

[66] T. T. Nguyen, N. Adams, D. Huang, M. M. Glymour, A. M. Allen, and Q. C. Nguyen, “The association between state-level racial attitudes assessed from twitter data and adverse birth outcomes: Observational study,” *JMIR public health and surveillance*, vol. 6, no. 3, p. e17103, 2020.

[67] M. Galesic, D. Barkoczi, A. M. Berdahl, D. Biro, G. Carbone, I. Giannoccaro, R. L. Goldstone, C. Gonzalez, A. Kandler, A. B. Kao, *et al.*, “Beyond collective intelligence: Collective adaptation,” *Journal of the Royal Society interface*, vol. 20, no. 200, p. 20220736, 2023.

[68] D. Garcia, M. Galesic, and H. Olsson, “The psychology of collectives,” *Perspectives on Psychological Science*, vol. 19, no. 2, pp. 316–319, 2024.

[69] W. Dai, J. Tao, X. Yan, Z. Feng, and J. Chen, “Addressing unintended bias in toxicity detection: An lstm and attention-based approach,” in *2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pp. 375–379, IEEE, 2023.

[70] L. Oswald, “Effects of preemptive empathy interventions on reply toxicity among highly active social media users,” *Public Discourse in Online Environments*, 2023.

[71] Pew Research Center, “Teens and cyberbullying 2022,” December 2022.

[72] Pew Research Center, “The state of online harassment,” January 2021.

[73] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. Kelley, D. Kumar, *et al.*, “Sok: Hate, harassment, and the changing landscape of online abuse,” in *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 247–267, IEEE, 2021.

[74] C. A. Bail, *Breaking the social media prism: How to make our platforms less polarizing*. Princeton: Princeton University Press, 2022.

[75] A. A. Anderson, D. Brossard, D. A. Scheufele, M. A. Xenos, and P. Ladwig, “The “nasty effect:” online incivility and risk perceptions of emerging technologies,” *Journal of computer-mediated communication*, vol. 19, no. 3, pp. 373–387, 2014.

[76] J. Kim, A. Guess, B. Nyhan, and J. Reifler, “The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity,” *Journal of Communication*, vol. 71, no. 6, pp. 922–946, 2021.

[77] A. Bor and M. B. Petersen, “The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis,” *American political science review*, vol. 16, no. 1, pp. 1–18, 2022.

[78] H. Hosseinmardi, A. Ghasemian, A. Clauset, M. Mobius, D. Rothschild, and D. Watts, “Examining the consumption of radical content on youtube,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 32, p. e2101967118, 2021.

[79] M. Haroon, M. Wojcieszak, A. Chhabra, X. Liu, P. Mohapatra, and Z. Shafiq, “Auditing youtube’s recommendation system for ideologically congenial, extreme, and problematic recommendations,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 50, p. e2213020120, 2023.

[80] M. Avalle, N. Di Marco, G. Etta, E. Sangiorgio, S. Alipour, A. Bonetti, L. Alvisi, A. Scala, A. Baronchelli, M. Cinelli, *et al.*, “Persistent interaction patterns across social media platforms and over time,” *Nature*, vol. 628, no. 8008, pp. 582–589, 2024.

[81] E. Wulczyn, N. Thain, and L. Dixon, “Ex machina: Personal attacks seen at scale,” in *Proceedings of the 26th international conference on world wide web*, pp. 1391–1399, 2017.

[82] A. Lees, V. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman, “A new generation of perspective api: Efficient multilingual character-level transformers,” in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 3197–3207, 2022.

[83] “Perspective api - how it works.” https://www.perspectiveapi.com/how-it-works/. Accessed: 21 October 2024.

[84] B. Vidgen and L. Derczynski, “Directions in abusive language training data, a systematic review: Garbage in, garbage out,” *Plos one*, vol. 15, no. 12, p. e0243300, 2020.

[85] Jigsaw, “Unintended Bias and Identity Terms,” *Medium*, 9 March 2018. https://medium.com/jigsaw/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23.

[86] G. Nogara, F. Pierri, S. Cresci, L. Luceri, P. Törnberg, and S. Giordano, “Toxic bias: Perspective api misreads german as more toxic,” 2023. Preprint at https://arxiv.org/abs/2312.12651.

[87] A. Sheth, V. L. Shalin, and U. Kursuncu, “Defining and detecting toxicity on social media: context and knowledge are key,” *Neurocomputing*, vol. 490, pp. 312–318, 2022.

[88] A. Rajadesingan, P. Resnick, and C. Budak, “Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits,” in *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 557–568, 2020.

[89] T. Hopp, C. J. Vargo, L. Dixon, and N. Thain, “Correlating self-report and trace data measures of incivility: A proof of concept,” *Social Science Computer Review*, vol. 38, no. 5, pp. 584–599, 2020.

[90] M. T. Thai, W. Wu, and H. Xiong, *Big data in complex and social networks*. CRC Press, 2016.

[91] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[92] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[93] S. J. Frenda, E. D. Knowles, W. Saletan, and E. F. Loftus, “False memories of fabricated political events,” *Journal of Experimental Social Psychology*, vol. 49, no. 2, pp. 280–286, 2013.

[94] G. Murphy, E. F. Loftus, R. H. Grady, L. J. Levine, and C. M. Greene, “False memories for fake news during ireland’s abortion referendum,” *Psychological science*, vol. 30, no. 10, pp. 1449–1459, 2019.

[95] W. Wei and X. Wan, “Learning to identify ambiguous and misleading news headlines,” *arXiv preprint arXiv:1705.06031*, 2017.

[96] U. K. Ecker, S. Lewandowsky, E. P. Chang, and R. Pillai, “The effects of subtle misinformation in news headlines.,” *Journal of experimental psychology: applied*, vol. 20, no. 4, p. 323, 2014.

[97] M. Cha, F. Benevenuto, H. Haddadi, and K. Gummadi, “The world of connections and information flow in twitter,” *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 42, no. 4, pp. 991–998, 2012.

[98] J. Allen, A. A. Arechar, G. Pennycook, and D. G. Rand, “Scaling up fact-checking using the wisdom of crowds,” *Science advances*, vol. 7, no. 36, p. eabf4393, 2021.

[99] N. Hassan, B. Adair, J. T. Hamilton, C. Li, M. Tremayne, J. Yang, and C. Yu, “The quest to automate fact-checking,” in *Proceedings of the 2015 computation+ journalism symposium*, Citeseer, 2015.

[100] T. Elsayed, P. Nakov, A. Barrón-Cedeno, M. Hasanain, R. Suwaileh, G. Da San Martino, and P. Atanasova, “Overview of the clef-2019 checkthat! lab: automatic identification and verification of claims,” in *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pp. 301–321, Springer, 2019.

[101] E. Estellés-Arolas and F. González-Ladrón-de Guevara, “Towards an integrated crowdsourcing definition,” *Journal of Information science*, vol. 38, no. 2, pp. 189–200, 2012.

[102] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, “Evidence for a collective intelligence factor in the performance of human groups,” *Sciences*, vol. 330, no. 6004, pp. 686–688, 2010.

[103] C. Riedl, Y. J. Kim, P. Gupta, T. W. Malone, and A. W. Woolley, “Quantifying collective intelligence in human groups,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 21, p. e2005737118, 2021.

[104] F. Gaisbauer, E. Olbrich, and S. Banisch, "Dynamics of opinion expression," *Physical Review E*, vol. 102, no. 4, p. 042303, 2020.

[105] N. Persily and J. A. Tucker, "Introduction," in *Social media and democracy: The state of the field, prospects for reform* (N. Persily and J. A. Tucker, eds.), Cambridge: Cambridge University Press, 2020.

[106] R. Huckfeldt, J. J. Mondak, M. Hayes, M. T. Pietryka, and J. Reilly, "Networks, interdependence, and social influence in politics," in *The Oxford Handbook of Political Psychology* (L. Huddy, D. Sears, and J. Levy, eds.), Oxford: Oxford University Press, 09 2013.

[107] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Reviews of modern physics*, vol. 81, no. 2, pp. 591–646, 2009.

[108] A. F. Hayes and J. Matthes, "Self-censorship, the spiral of silence, and contemporary political communication," in *The Oxford Handbook of Political Communication* (K. Kenski and K. Jamieson, eds.), Oxford: Oxford University Press, 2014.

[109] P. Porten-Cheé and C. Eilders, "Spiral of silence online: How online communication affects opinion climate perception and opinion expression regarding the climate change debate," *Studies in communication sciences*, vol. 15, no. 1, pp. 143–150, 2015.

[110] S. S. Ho and D. M. McLeod, "Social-psychological influences on opinion expression in face-to-face and computer-mediated communication," *Communication research*, vol. 35, no. 2, pp. 190–207, 2008.

[111] J. Matthes, K. Rios Morrison, and C. Schemer, "A spiral of silence for some: Attitude certainty and the expression of political minority opinions," *Communication Research*, vol. 37, no. 6, pp. 774–800, 2010.

[112] E. Folador, S. Tiwari, C. Da Paz Barbosa, S. Jamal, M. Da Costa Schulze, D. Barh, and V. Azevedo, "Protein-protein interactions: an overview," in *Encyclopedia of Bioinformatics and Computational Biology* (S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, eds.), pp. 821–833, Oxford: Academic Press, 2019.

[113] D. Ramage, "Cs229: Hidden markov models fundamentals," December 2007. https://cs229.stanford.edu/section/cs229-hmm.pdf.

[114] S. Istrail, "Hmm: The learning problem," March 2020. https://cs.brown.edu/courses/csci1820/resources/HMMs_The_Learning_Problem_slides.pdf.

[115] J. Eisner, "An interactive spreadsheet for teaching the forward-backward algorithm," in *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching*

natural language processing and computational linguistics, pp. 10–18, 2002. https://aclanthology.org/W02-0102.pdf.

[116] J. Mastrine, "Introducing the AllSides Media Bias Chart," *AllSides*, 1 February 2019. https://www.allsides.com/blog/introducing-allsides-media-bias-chart.

[117] AllSides Staff, "Introducing the AllSides Media Bias Meter™," *AllSides*, 3 January 2023. https://www.allsides.com/blog/introducing-allsides-media-bias-meter.

[118] "Perspective api - attributes & languages." https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US. Accessed: 21 October 2024.

[119] "Perspective api - score." https://developers.perspectiveapi.com/s/about-the-api-score?language=en_US. Accessed: 21 October 2024.

[120] M. Saveski, B. Roy, and D. Roy, "The structure of toxic conversations on twitter," in *Proceedings of the web conference 2021* (Leskovec, J., *et al*, ed.), (New YorkNYUnited States), pp. 1086–1097, Association for Computing Machinery, 2021.

[121] W. Brady, J. Wills, J. Jost, J. Tucker, and J. Van Bavel, "Emotion shapes the diffusion of moralized content in social networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 28, pp. 7313–7318, 2017.

[122] E. McLaughlin, "Anger erupts in American cities after charging decision in Breonna Taylor case," *CNN*, 28 September 2020. https://edition.cnn.com/2020/09/28/us/weekend-protests-breonna-taylor/index.html.

[123] BBC Staff, "George Floyd: Timeline of black deaths and protests," *BBC*, 22 April 2021. https://www.bbc.com/news/world-us-canada-52905408.

[124] P. Huang, "'A Loss To The Whole Society': U.S. COVID-19 Death Toll Reaches 500,000," *NPR*, 22 February 2021. https://www.npr.org/sections/health-shots/2021/02/22/969494791/a-loss-to-the-whole-society-u-s-covid-19-death-toll-reaches-500-000.

[125] I. Vayansky and S. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, p. 101582, 2020.

[126] M. Gerlach, T. P. Peixoto, and E. G. Altmann, "A network approach to topic models," *Science advances*, vol. 4, no. 7, p. eaaq1360, 2018.

[127] C. Blex and T. Yasseri, “Positive algorithmic bias cannot stop fragmentation in homophilic networks,” *The Journal of Mathematical Sociology*, vol. 46, no. 1, pp. 80–97, 2022.

[128] M. Falkenberg, F. Zollo, W. Quattrociocchi, J. Pfeffer, and A. Baronchelli, “Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries,” *Nature Communications*, vol. 15, no. 1, p. 9560, 2024.

[129] L. Oswald, “More than news! mapping the deliberative potential of a political online ecosystem with digital trace data,” *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1–16, 2024.

[130] “X.com community notes guide.” Accessed: April 3, 2024.

[131] J. Allen, C. Martel, and D. G. Rand, “Birds of a feather don’t fact-check each other: Partisanship and the evaluation of news in twitter’s birdwatch crowdsourced fact-checking program,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022.

[132] T. W. Malone, R. Laubacher, and C. Dellarocas, “The collective intelligence genome,” *MIT Sloan management review*, 2010.

[133] A. Moore, “Democratic reason: Politics, collective intelligence and the rule of the many,” *Contemporary Political Theory*, vol. 13, pp. e12–e15, 2014.

[134] I. Mergel and K. C. Desouza, “Implementing open innovation in the public sector: The case of challenge. gov,” *Public administration review*, vol. 73, no. 6, pp. 882–890, 2013.

[135] F. Shi, M. Teplitskiy, E. Duede, and J. A. Evans, “The wisdom of polarized crowds,” *Nature human behaviour*, vol. 3, no. 4, pp. 329–336, 2019.

[136] G. Pennycook and D. G. Rand, “Fighting misinformation on social media using crowdsourced judgments of news source quality,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 7, pp. 2521–2526, 2019.

[137] O. Arazy, O. Nov, R. Patterson, and L. Yeo, “Information quality in wikipedia: The effects of group composition and task conflict,” *Journal of management information systems*, vol. 27, no. 4, pp. 71–98, 2011.

[138] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész, “Dynamics of conflicts in wikipedia,” *PloS one*, vol. 7, no. 6, p. e38869, 2012.

[139] D. Guilbeault, J. Becker, and D. Centola, “Social learning and partisan bias in the interpretation of climate trends,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 39, pp. 9714–9719, 2018.

[140] A. Combs, G. Tierney, B. Guay, F. Merhout, C. Bail, D. S. Hillygus, and A. Volfovsky, "Anonymous cross-party conversations can decrease political polarization: A field experiment on a mobile chat platform," *Nature human behaviour*, vol. 7, no. 9, p. 1454–1461, 2023.

[141] T. Yasseri and F. Menczer, "Can crowdsourcing rescue the social marketplace of ideas?," *Communications of the ACM*, vol. 66, no. 9, pp. 42–45, 2023.

[142] E. Danchin, L.-A. Giraldeau, T. J. Valone, and R. H. Wagner, "Public information: from nosy neighbors to cultural evolution," *Science*, vol. 305, no. 5683, pp. 487–491, 2004.

[143] W. B. G. Jarvis and R. E. Petty, "The need to evaluate," *Journal of personality and social psychology*, vol. 70, no. 1, p. 172, 1996.

[144] J. Dalege, D. Borsboom, F. van Harreveld, and H. L. van der Maas, "A network perspective on attitude strength: Testing the connectivity hypothesis," *Social Psychological and Personality Science*, vol. 10, no. 6, pp. 746–756, 2019.

[145] J. Hox, M. Moerbeek, and R. Van de Schoot, *Multilevel analysis: Techniques and applications*. New York: Routledge, 2017.

[146] K. L. McLoughlin, W. J. Brady, A. Goolsbee, B. Kaiser, K. Klonick, and M. Crockett, "Misinformation exploits outrage to spread online," *Science*, vol. 386, no. 6725, pp. 991–996, 2024.

[147] M. Duncan, A. Pelled, D. Wise, S. Ghosh, Y. Shan, M. Zheng, and D. McLeod, "Staying silent and speaking out in online comment sections: The influence of spiral of silence and corrective action in reaction to news," *Computers in Human Behavior*, vol. 102, pp. 192–205, 2020.

[148] H. Jo, M. Karsai, J. Kertész, and K. Kaski, "Circadian pattern and burstiness in mobile phone communication," *New Journal of Physics*, vol. 14, no. 1, p. 013055, 2012.

Appendix A

Supplementary Material for Chapter 2

A.1 Data description

Table A.1: **Descriptive Statistics by Comment Type.** The dataset contains top-level comments and replies from videos posted by six prominent U.S. news outlets chosen for their active YouTube comment sections and for representing a broad spectrum of political ideologies. The time frame begins in September 2020, and extends through April 2021.

	Comments								
	Channel	Top Level	%[1]	Toxicity[2]	Insult[3]	Replies	%	Toxicity	Insult
Lean left	ABC News	1993690	49.8	20.3	20.8	2009172	50.2	17.1	18.1
	CBS	1443344	49.1	6.5	6.8	1496646	50.9	4.8	5.4
	CNN	5945976	59.0	12.0	11.0	4128018	41.0	7.6	7.3
Right	OAN	403054	65.8	6.7	6.0	209545	34.2	5.5	5.7
	Newsmax	1881808	66.4	7.0	7.9	950237	33.6	7.0	8.2
	Fox News	7387795	62.9	10.0	10.5	4355601	37.1	8.0	8.9
	ALL	19055667	59.2	11.1	11.1	13149219	40.8	8.8	9.3

1. As a percentage of all comments in channel

2. Percentage of toxic comments by comment type. A comment is toxic if its toxicity score is greater than 0.6.

3. Percentage of insulting comments by comment type. A comment is insulting if its insult score is greater than 0.6.

Figure A.1 presents distribution functions for key metrics in the YouTube dataset. Panel (a) shows the distribution of interevent times, measured in seconds, for commenting behavior. Reflecting typical human communication patterns, which are often bursty and marked by irregular interaction intervals [148], our dataset shows both short and long interevent times, indicating a bursty commenting frequency rather than a steady interaction rate. Panel (b) shows the distribution of conversation length, measured by the number of comments per conversation. Generally, conversations have between 2 and 500 comments, with only a few sustaining high engagement, while the majority remain relatively brief. Panel (c) shows the number of comments per user, an indicator of user activity level, while panel (d) shows the number of comments per video, reflecting overall engagement levels. Both metrics reveal similar heterogeneous patterns: a small number of users are highly active, while most post only occasionally, and specific videos gain popularity, attracting a large volume of comments.

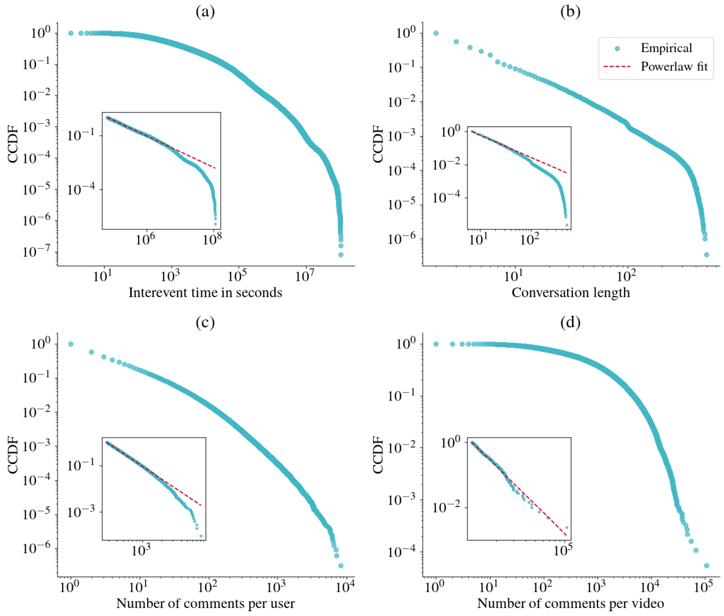


Figure A.1: **Complementary Cumulative Distribution Functions (CCDF) for engagement metrics in YouTube.** Figures show empirical data (blue dots) and power law fits (red dashed lines) for each plot. Insets highlight alignment with power law distributions at different scales. **(a)** Distribution of interevent times (in seconds) for user commenting behavior. The distribution is heterogeneous, as evidenced by the heavy-tailed CCDF. The presence of both short and long interevent times suggests burstiness in commenting frequency, rather than a consistent interaction rate. **(b)** Distribution of conversation length, measured by the number of comments per conversation. This distribution follows a heavy-tailed pattern, indicating that while most conversations are brief, a few are lengthy and maintain high engagement levels. **(c)** Distribution of the number of comments per user, illustrating user activity levels. The CCDF shows a heterogeneous and heavy-tailed distribution, with a few users being highly active while the majority post occasionally. **(d)** CCDF of the number of comments per video, illustrating overall engagement levels per video. This distribution underscores the nature of content virality in online platforms, where specific videos become popular and attract a significant number of comments.

A.2 Bias ratings

In January 2023, AllSides launched the AllSides Media Bias Meter. Before this update, the website classified media outlets into five bias categories: Left, Lean Left, Center, Lean Right,

Table A.2: **AllSides Media Bias Ratings.** The AllSides Media Bias Meter provides a nuanced rating of media outlets on a scale from -6 (farthest Left) to +6 (farthest Right). The bias meter rating is based on a combination of methods, including Editorial Reviews conducted by a multipartisan panel and a Blind Bias Survey where respondents rate outlets on an 11-point scale. The final rating typically averages results from these methods, with adjustments made for data consistency and recency, and more recent reviews carrying greater weight.

Channel	Bias Meter Rating	Bias Meter Score	Rating Confidence
ABC[1]	Lean Left	-2,40	High
CBS[2]	Lean Left	-1,50	High
CNN[3]	Lean Left	-1,30	High
OAN[4]	Right	3,10	Medium
Newsmax[5]	Right	3,28	Low/Initial
Fox News[6]	Right	3,88	Medium

1. https://www.allsides.com/news-source/abc-news-media-bias Retrieved August 22, 2024

2. https://www.allsides.com/news-source/cbs-news-media-bias Retrieved August 22, 2024

3. https://www.allsides.com/news-source/cnn-media-bias Retrieved August 22, 2024

4. https://www.allsides.com/news-source/one-america-news-network-media-bias Retrieved August 22, 2024

5. https://www.allsides.com/news-source/newsmax Retrieved August 22, 2024

6. https://www.allsides.com/news-source/fox-news-media-bias Retrieved August 22, 2024

and Right. The new bias meter provides a more detailed assessment, rating outlets on a scale from -6 to +6, with 0.0 representing a perfect Center, -6 indicating the farthest Left, and +6 indicating the farthest Right.

The bias meter rating is derived from a combination of methods. First, AllSides conducts Editorial Reviews with a multipartisan panel representing different political views. Each panelist reviews news content individually, identifies bias indicators, and provides their initial rating. The panel then discusses their findings, and members may revise their ratings. The ratings are averaged by grouping them into three categories (Left, Center, Right) and calculating a weighted average. If all panelists agree that the final rating should differ from the calculated average, they can unanimously override it.

In addition to Editorial Reviews, AllSides conducts a Blind Bias Survey to assess bias meter ratings. Respondents from different political perspectives rate headlines and news reports on an 11-point Likert scale. AllSides then calculates an average score for each bias group and an overall weighted average to produce the Blind Bias Survey result. This result uses a scale of -9 to +9, which differs from the AllSides Media Bias Meter scale of -6 to +6. Despite the possibility of higher scores in the survey, AllSides caps the final bias rating at -6 and +6 to align with the methodology of other analyses like Editorial Reviews.

The Final Bias Meter rating is determined by the number and timing of review methods applied to a source. If multiple reviews have been conducted, AllSides considers all data and assigns an overall numerical value, usually averaging results from Editorial Reviews and Blind Bias Surveys. Adjustments may be made based on factors like data consistency and recency, with full documentation of any changes. If only one methodology is available, the rating will reflect that result. When multiple methods were applied at different times, more recent reviews are given greater weight in calculating the final rating.

A.3 Perspective API prevalence all attributes

A.4 Hidden Markov model fit details

A Hidden Markov model (HMM) is a statistical framework designed to analyze systems where observed data is generated by underlying hidden states that are not directly visible. As a type of dynamic Bayesian network, HMMs excel in modeling temporal and sequential data, making them versatile for applications such as speech recognition, biological sequence analysis, and handwriting recognition. HMMs operate on the principle of the Markov process, where the future state depends only on the present state, not on the sequence of past states. However, unlike traditional Markov models, HMMs incorporate hidden states, which are inferred through

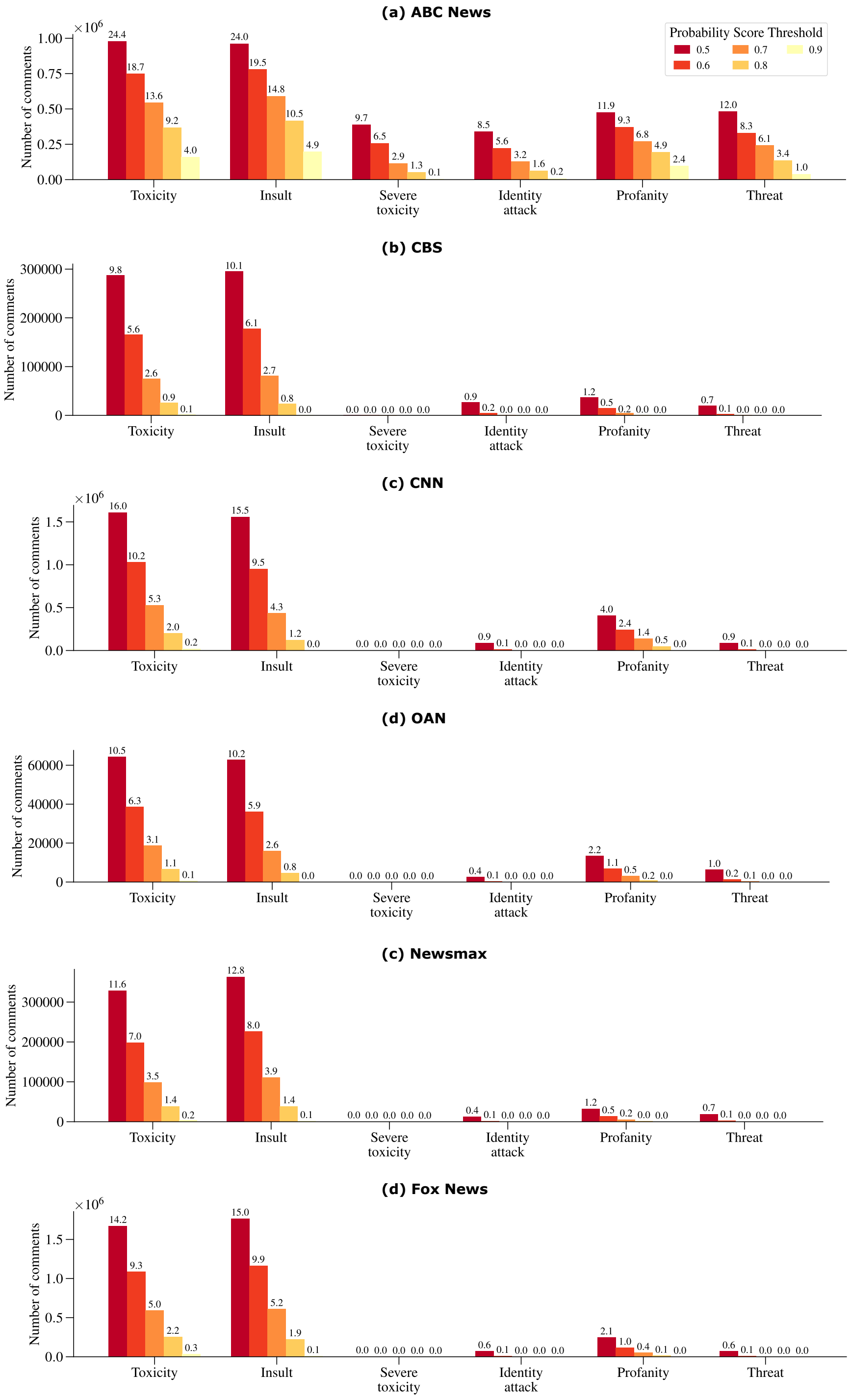


Figure A.2: Prevalence of six emotions by channel in the dataset. The height of the bars indicates the absolute frequency of each emotion, with percentages shown at the top. Bar colors represent different probability score thresholds.

observed events or symbols. The model relies on transition probabilities, which define the likelihood of moving between hidden states, and emission probabilities, which describe the likelihood of observing specific symbols given a state.

In this study, we utilize HMMs to analyze the relationship between negative sentiment and disengagement in YouTube comment threads. Our approach involves tailoring HMM parameters, such as transition and emission probabilities, to align with conversation sequences from six different channels. We categorize videos in two ways: by their news media channel publishers and by their topics, as identified through topic modeling. To ensure robustness, we use alternative group definitions, confirming that the observed patterns are not dependent on specific ensemble classifications. Conversations are modeled as sequences of observable states (X_2 for non-toxic comments, X_3 for toxic ones) with an appended zero ($X_1 = 0$) marking the conversation's end. This transformation enables a structured analysis of comment sequences, revealing hidden conversational dynamics.

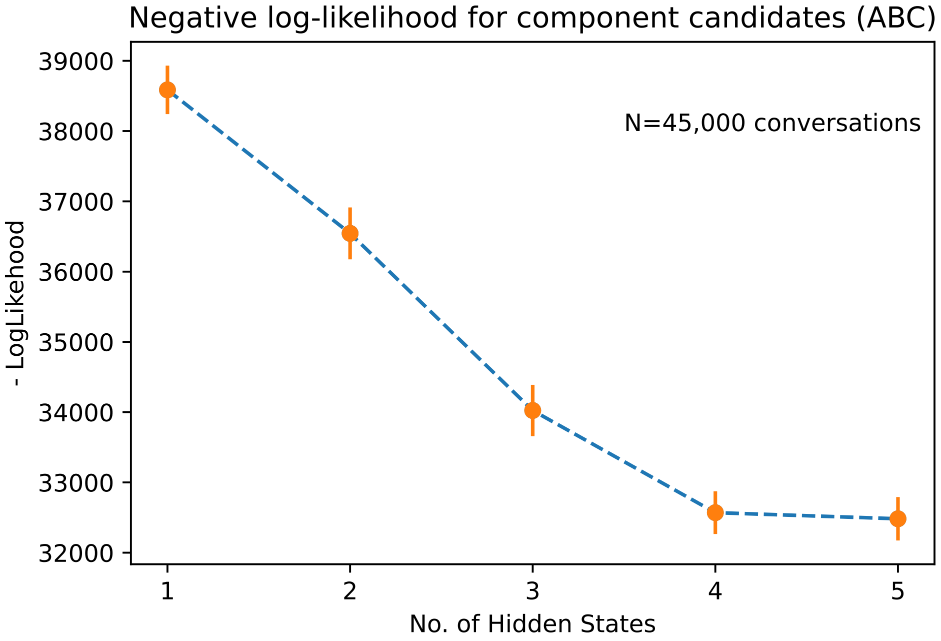
Given the dataset's size, we employed a sampling process to fit the model parameters. Each realization involved randomly selecting 45,000 conversations, split into training (80%) and test (20%) sets. This iterative process, requiring approximately two hours per realization, allowed for efficient parameter estimation across multiple subsets of the data (see Table A.3 for number of fits). Although we fixed the number of hidden states at two, robustness checks with three- to five-state models revealed that the four-state model achieved the highest log-likelihood (see Figure A.3). However, qualitative analysis showed that the additional states primarily refined the characterization of one state (Z_2), associated with the tone of conversations preceding disengagement. While informative, this refinement extends beyond the scope of our study, which focuses on distinguishing between self-censorship and active engagement. Consistent with prior research, we selected the simpler two-state model to avoid overfitting, as model selection procedures for HMMs often favor overly complex models.

Table A.3: Number of Hidden Markov model fits

	Channel	Threads		No. Fits	
		N	%	Toxicity	Insult
Lean left	ABC News	447756	15.5	4139	3993
	CBS	356538	12.4	4179	3443
	CNN	888010	30.8	4758	4391
Right	OAN	48318	1.7	4680	4703
	Newsmax	187440	6.5	3698	3817
	Fox News	955409	33.1	3410	3248
	ALL	2883471	100.0		

A.5 Hidden Markov model fit robustness checks

Figure A.3: Negative log-likelihood for component candidates, for conversations on ABC news



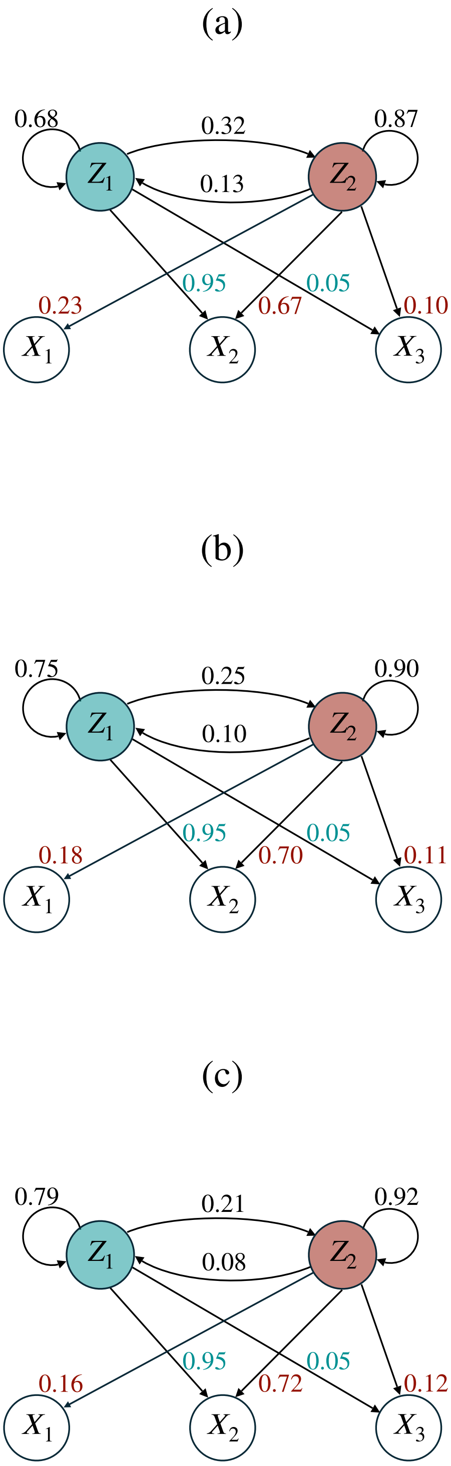


Figure A.4: Inferred transition and emission probabilities for a 2-state Hidden Markov Model, filtering short conversations in increasing order. Panel (a) shows results after excluding conversations of length 2. Panel (b) presents results after excluding conversations of length 3 or shorter. Panel (c) displays results after excluding conversations of length 4 or shorter. The shown probabilities are averages over multiple model fits, with standard errors consistently below 0.001. The model results remain consistent across all panels, indicating that filtering short conversations has no significant impact on the overall inferred probabilities.

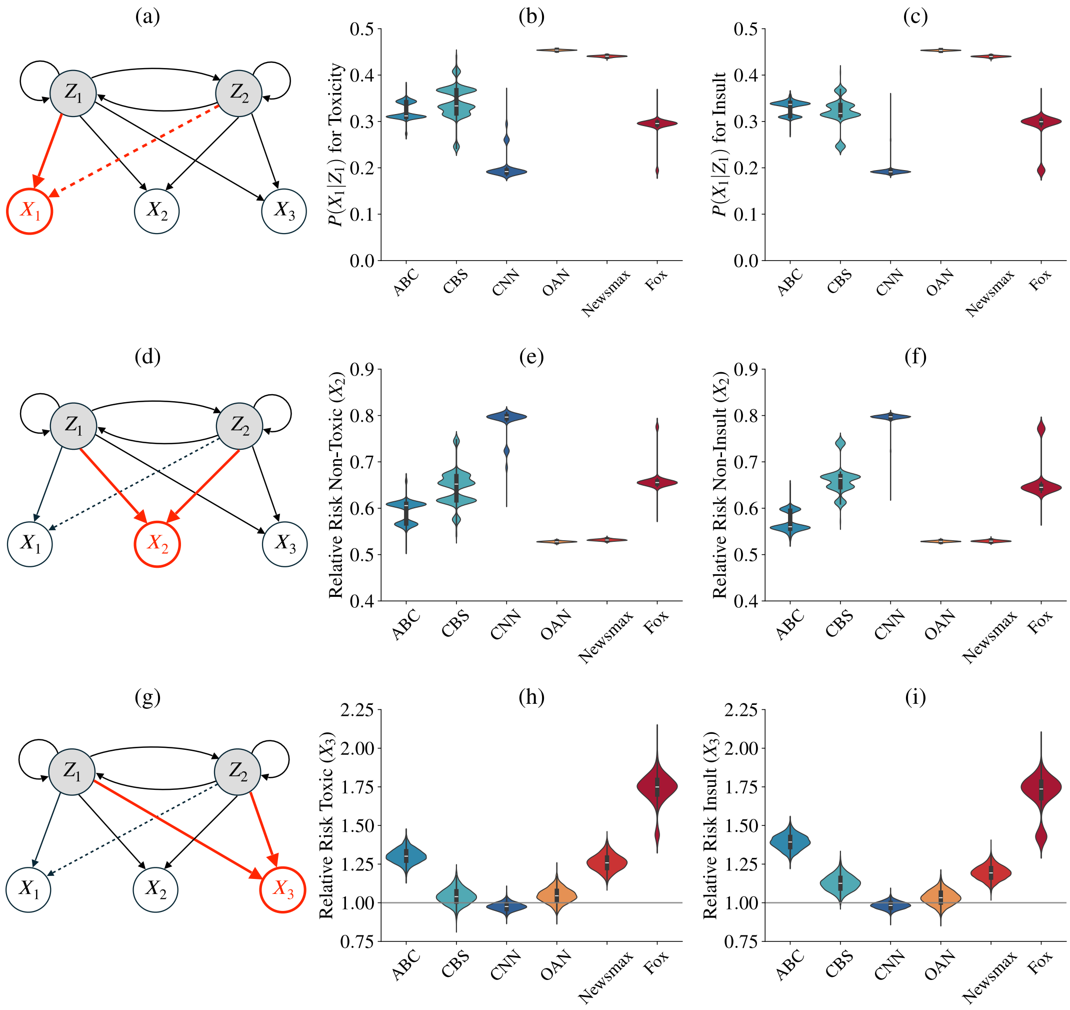


Figure A.5: **Inferred emission probabilities and relative risks by news channels.** Panels **(a)-(c)** show emission probabilities associated with X_1, with panel **(b)** detailing $P(X_1|Z_1)$ for toxic content, panel **(c)** for insulting content, and panel **(a)** presents these probabilities in a diagrammatic way. Notably, $P(X_1|Z_2) = 0$ across all channels (indicated by a dotted arrow in panels **(a)**, **(d)**, and **(g)**), suggesting that conversations do not conclude in state Z_2. This, combined with $P(X_1|Z_1) > 0$, identifies state Z_1 as the likely terminal state, while state Z_2 corresponds to earlier stages in a conversation. Panels **(d)-(f)** display relative risk findings for X_2, with non-toxic posts (panel **(e)**) and non-insulting posts (panel **(f)**). Here, $RR_{X_2} < 1$ across all channels, suggesting that non-toxic or non-insulting activity is more common in state Z_2. Panels **(g)-(i)** summarize the relative risk findings for X_3, with toxic posts in panel **(h)** and insulting posts in panel **(i)**. In almost all channels (with CNN as an exception), $RR_{X_3} > 1$, indicating that toxic or insulting posts are more likely when a conversation is in state Z_1. This effect is most pronounced for Fox News (a right-leaning channel), while among left-leaning channels, ABC News exhibits the highest relative risk for X_3.

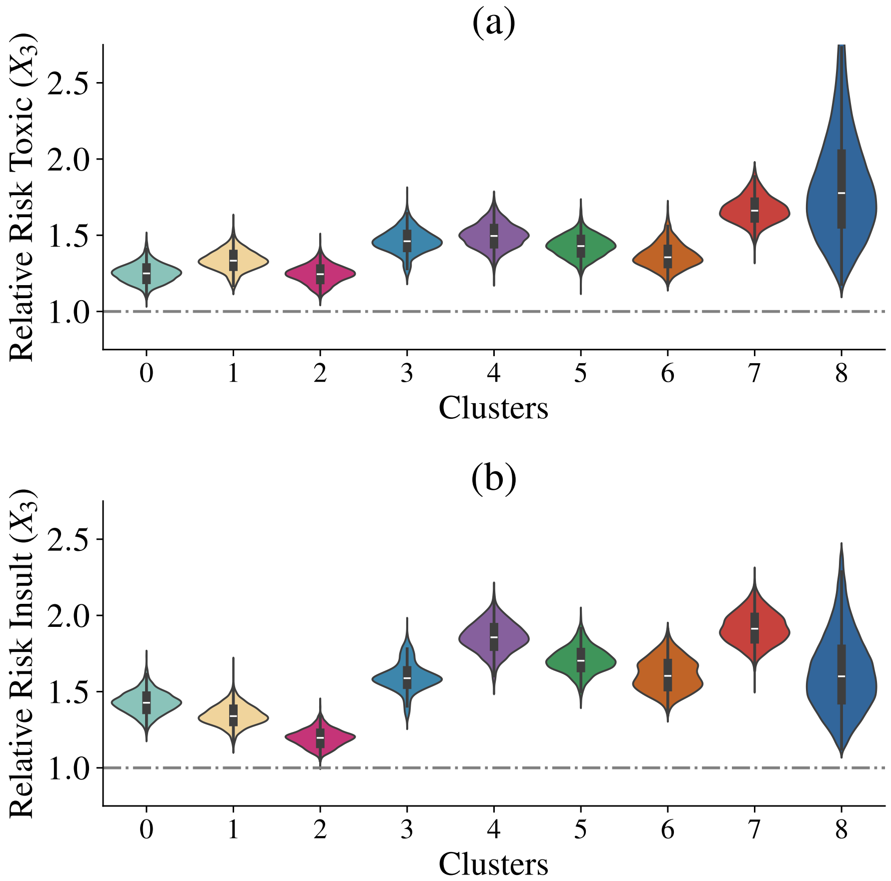


Figure A.6: **Inferred relative risks for toxic and insulting content by topic clusters. Clusters 1, 2, and 0:** Clusters focusing on the 2020 U.S. Presidential Election, January 6 Capitol Attack, and COVID-19 media coverage have low relative risks for toxic content (1.2–1.3) and insults, with cluster 0 slightly higher at just below $RR_{X_3} = 1.5$. **Clusters 3 and 6:** Clusters discussing the Election and Capitol Attack, along with additional topics like migration, societal inequality, and COVID-19 communication, show low relative risks for toxic content but higher risks for insults when debates reach terminal stages, exceeding $RR_{X_3} = 1.5$. **Clusters 5 and 8:** Clusters focusing on COVID-19 vaccination show contrasting behavior, where cluster 8 has the highest relative risk for toxic content, likely due to a small sample size with over representation of OAN videos, while cluster 5 has high insult risks exceeding $RR_{X_3} = 1.5$, reflecting the controversial nature of vaccination discussions. **Clusters 4 and 7:** Videos addressing police brutality and the Black Lives Matter movement exhibit elevated risks for insults, with cluster 7 also showing heightened toxicity, underscoring the emotionally charged and contentious nature of these topics, particularly in the final stages of discussions.

Appendix B

Supplementary Material for Chapter 3

B.1 Quantitative analysis of the notes

To gain a more comprehensive understanding of the notes, we extracted several quantitative metrics. The metrics are defined as follows:

- **Length**: The length of the note, measured by the number of words.
- **Links**: The number of hyperlinks included in the note as references.
- **Sophistication Level**: A standard score indicating readability, complexity, and grade level (according to the U.S. education system). This score is calculated using the ‘textstat‘ library (version 0.7.3) in Python (version 3.11).
- **Stats**: A binary variable indicating the presence or absence of numerical data in the text.

The distributions of the metrics of each category are demonstrated in Figure B.2.

The improvement in each of these metrics was calculated using the following equation:

$$I_{\mathrm{X}} = X^{ab} - \frac{1}{2}\left(X^{a} + X^{b}\right),$$

where X represents any of the aforementioned metrics. X^{ab} denotes the metric for the note co-authored by teams a and b, while X^{a} and X^{b} refer to the metrics for the notes written individually by teams a and b, respectively.

B.2 Qualitative analysis of the notes

Within the crowdsourcing experiment, the participants were asked to evaluate the notes based on the following criteria.

- **It provides context**: Does this note provide sufficient context for the Tweet, equipping the reader with enough information?
- **It is fair**: Is this note fair, impartial, and unbiased in tone?
- **It is relevant**: Does the note directly address the issues raised in the Tweet, or is it mere speculation or personal opinion unrelated to the Tweet's claims?
- **It is readable**: Is this note well-written, free from typos and grammatical errors and can be easily understood?
- **It has the right sources**: Are there any sources cited in the note, and are they reliable and supportive of Tweet's argument?
- **Overall Helpfulness**: On a scale of 0 (least helpful) to 10 (most helpful), rate the note based on its overall quality and helpfulness.

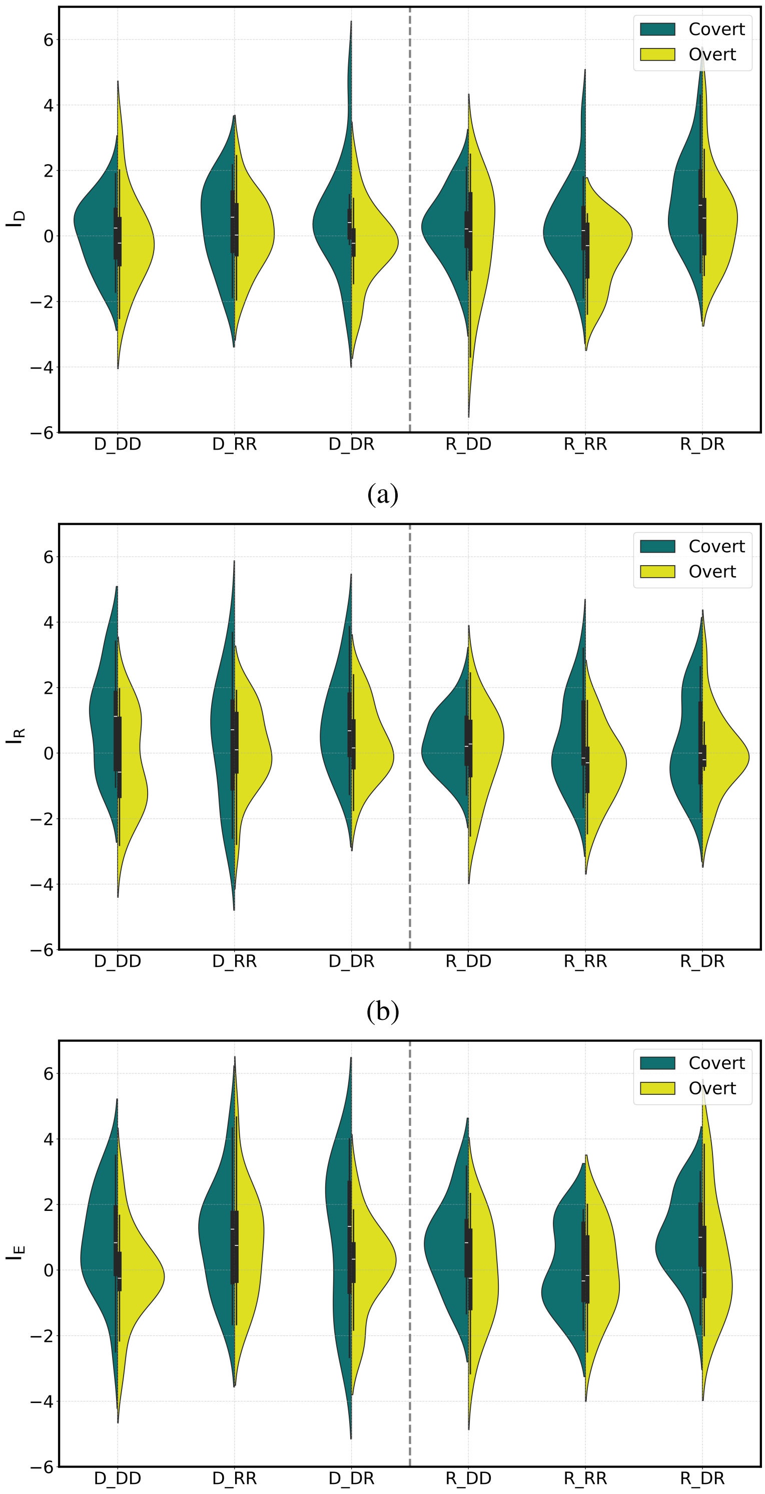
The helpfulness scores have been explained and analysed in length in the main text.

B.3 Bootstrap analysis

To examine the robustness of the results, we employed the following bootstrapping method. First, we randomly selected half of the tweets from both the Democratic and Republican datasets. Next, we categorized the data points based on team dynamics and treatments. We then calculated the mean of the helpfulness scores from each resulting distribution. This process was repeated 184,756 times, corresponding to the number of possible combinations when selecting 10 out of 20 options. The distribution of the means of each category is shown in figures B.3, B.4, and B.5.

Table B.1: Statistical comparison of the helpfulness scores between individuals and teams across different groups. The mean and standard deviation (in parentheses) are presented for each distribution, along with the t-value and p-value.

Group (N)	**Variable**	**Individual**	**Teams**	**t-value**	**p-value**
D_DD (N = 36)	H_D	5.1 (1.5)	5.1 (1.2)	-0.078	0.93
	H_R	5.1 (1.8)	4.8 (1.2)	0.98	0.33
	H_E	3.6 (1.9)	3.2 (1.5)	0.89	0.38
D_RR (N = 32)	H_D	4.1 (1.4)	4.0 (1.4)	0.62	0.54
	H_R	5.1 (1.4)	4.9 (1.3)	0.70	0.49
	H_E	3.5 (1.7)	2.6 (1.3)	2.2	0.028
D_RD (N = 35)	H_D	4.2 (1.5)	4.0 (1.2)	0.42	0.68
	H_R	4.9 (1.3)	4.3 (1.1)	2.0	0.05
	H_E	2.9 (1.7)	2.4 (1.1)	1.4	0.16
R_DD (N = 37)	H_D	5.6 (1.7)	5.1 (1.2)	1.2	0.24
	H_R	4.5 (1.4)	4.8 (1.2)	-0.92	0.36
	H_E	3.6 (1.2)	3.2 (1.5)	0.88	0.38
R_RR (N = 37)	H_D	4.2 (1.4)	3.9 (1.4)	0.98	0.33
	H_R	4.4 (1.3)	4.9 (1.3)	-1.7	0.095
	H_E	2.8 (1.7)	2.6 (1.3)	0.36	0.72
R_RD (N = 39)	H_D	6.0 (1.7)	4.0 (1.2)	5.7	2.2×10^{-7}
	H_R	4.9 (1.4)	4.3 (1.1)	2.0	0.050
	H_E	3.9 (2.0)	2.4 (1.1)	3.7	0.00035



(a)

(b)

(c)

Figure B.1: Improvement scores for notes in Covert v. Overt treatments by source of the evaluated tweet and group configuration. **Panel (a):** ratings from Democrats evaluators. **Panel (b):** ratings from Republicans evaluators. **Panel (c):** ratings from experts evaluators.

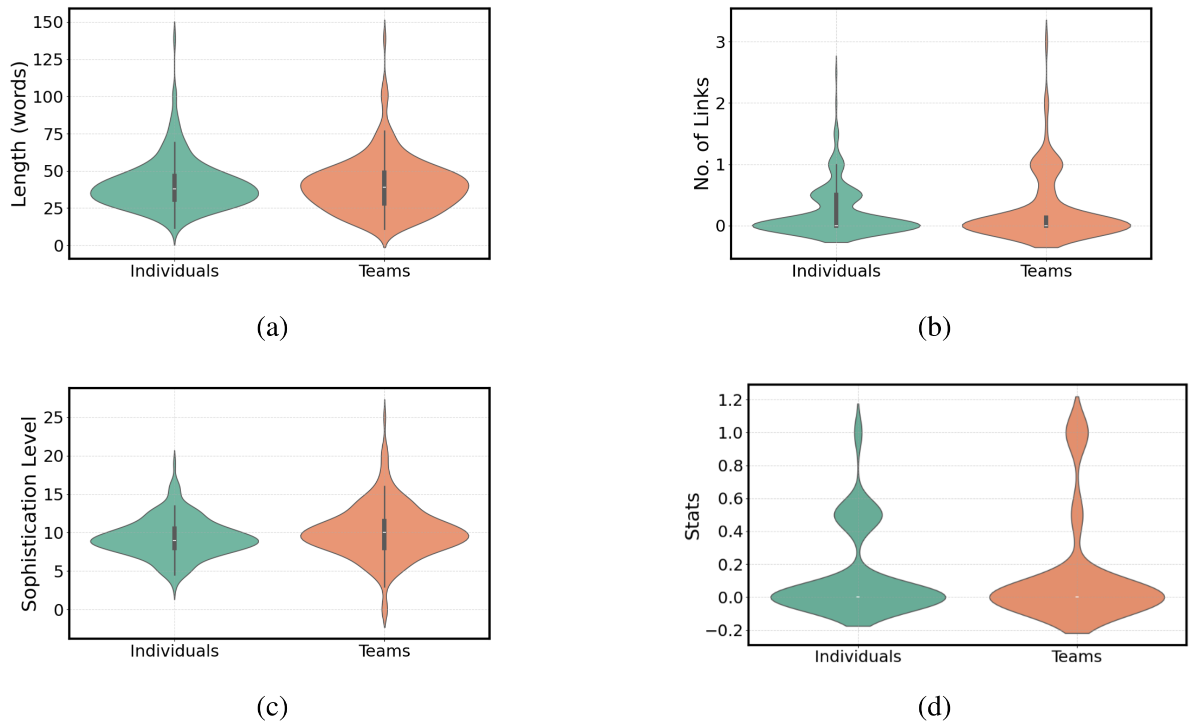
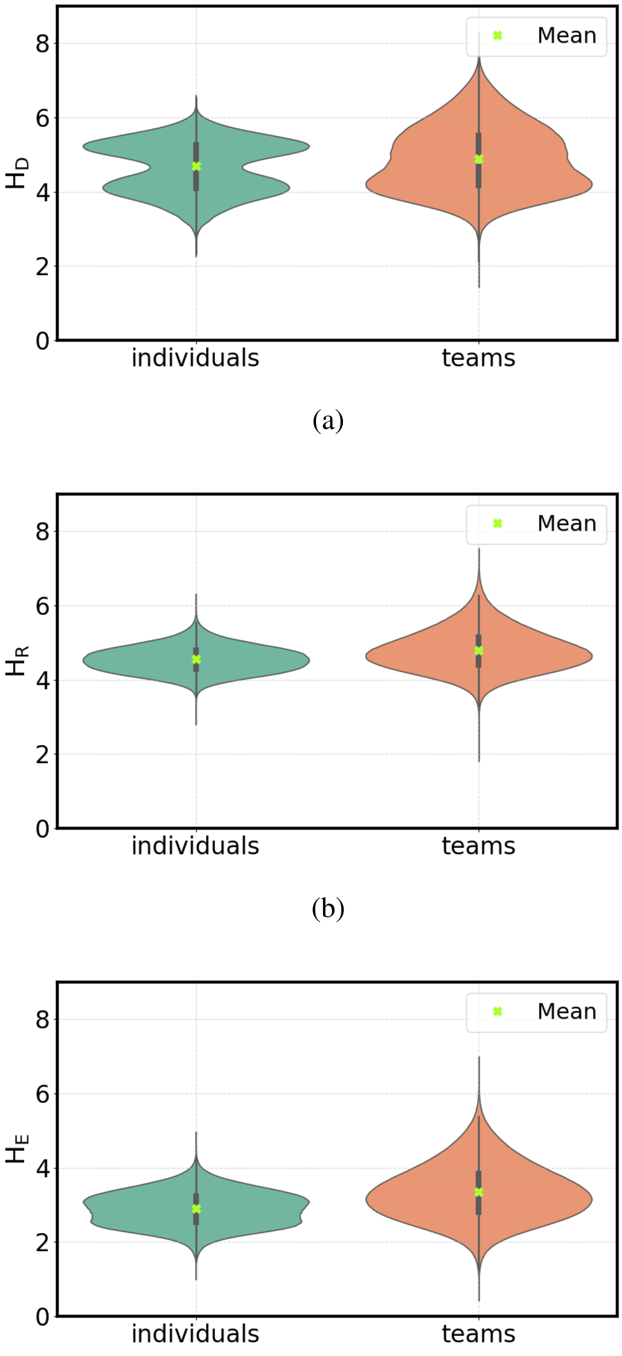


Figure B.2: The distribution of the quantitative measures for notes written by individuals and teams. **Panel (a):** Length of the notes by words. **Panel (b):** Number of the links used in the notes. **Panel (c):** The sophistication level of the notes. **Panel (d):** The likelihood of the existence of numbers in a note.

Table B.2: Comparison of Improvement of notes for notes written in the Covert and Overt treatment. The mean and standard deviation (in parentheses) are presented for each distribution, along with the t-value and p-value.

Group	**Variable**	**Covert**	**Overt**	**t-value**	**p-value**
D_DD (N = 36)	I_{D}	0.061 (1.0)	-0.11 (1.4)	0.43	0.67
	I_{R}	1.0 (1.5)	-0.34 (1.4)	2.8	0.0079
	I_{E}	0.88 (1.5)	-0.17 (1.3)	2.2	0.036
D_RR (N = 32)	I_{D}	0.33 (1.3)	0.12 (1.1)	0.48	0.64
	I_{R}	0.39 (1.9)	0.12 (1.2)	0.46	0.65
	I_{E}	0.98 (1.6)	0.78 (1.6)	0.34	0.73
D_RD (N = 39)	I_{D}	1.1 (1.4)	0.55 (1.4)	1.2	0.23
	I_{R}	0.29 (1.4)	-0.022 (1.2)	0.75	0.46
	I_{E}	0.99 (1.3)	0.42 (1.8)	1.1	0.27
R_DD (N = 37)	I_{D}	0.24 (1.0)	-0.062 (1.6)	0.67	0.51
	I_{R}	0.34 (0.90)	0.074 (1.3)	0.71	0.48
	I_{E}	0.75 (1.3)	-0.13 (1.5)	1.9	0.069
R_RR (N = 37)	I_{D}	0.33 (1.3)	-0.53 (0.99)	2.3	0.023
	I_{R}	0.35 (1.3)	-0.38 (1.1)	1.8	0.077
	I_{E}	0.079 (1.3)	-0.028 (1.34)	0.25	0.81
R_RD (N = 39)	I_{D}	1.1 (1.4)	0.55 (1.4)	1.2	0.23
	I_{R}	0.29 (1.4)	-0.022 (1.2)	0.75	0.46
	I_{E}	0.98 (1.3)	0.42 (1.8)	1.1	0.27

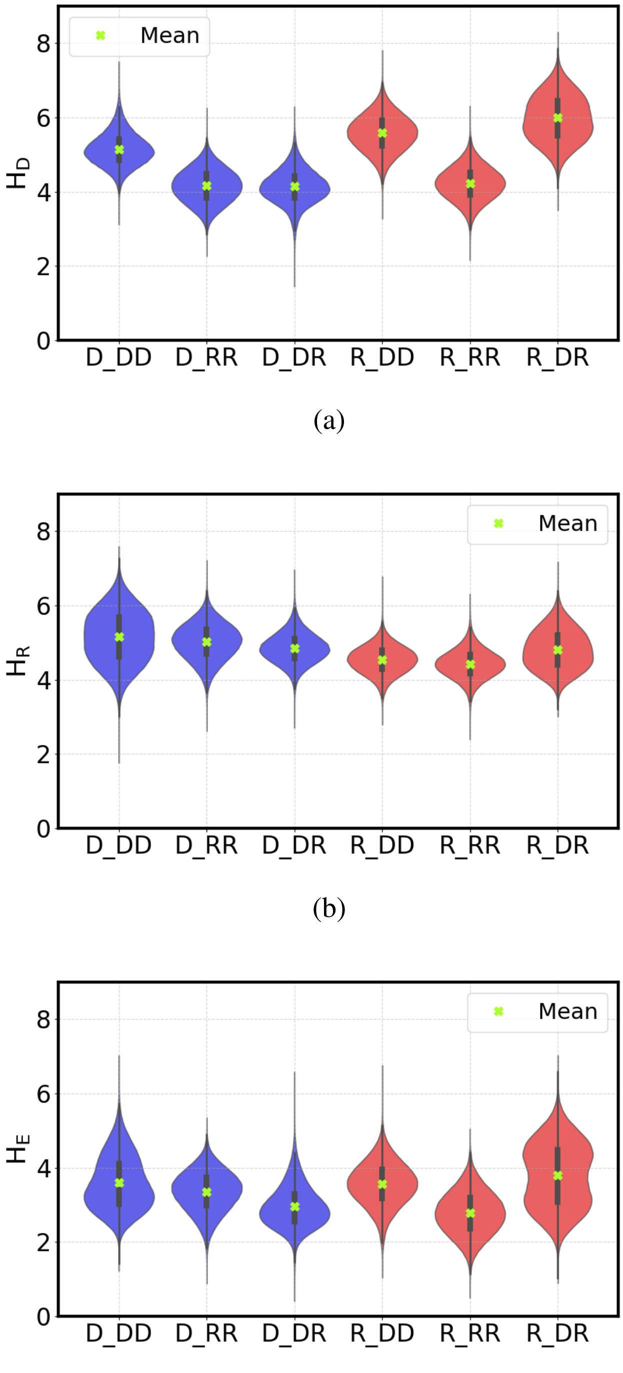


(a)

(b)

(c)

Figure B.3: Comparison of the distributions of helpfulness scores from the bootstrapping process in notes written by teams v. individuals. The mean of the original data points is marked by the green x. **Panel (a):** The helpfulness scores as evaluated by self-identified Democrats. **Panel (b):** The helpfulness scores as evaluated by self-identified Republicans. **Panel (c):** The helpfulness scores as evaluated by experts.

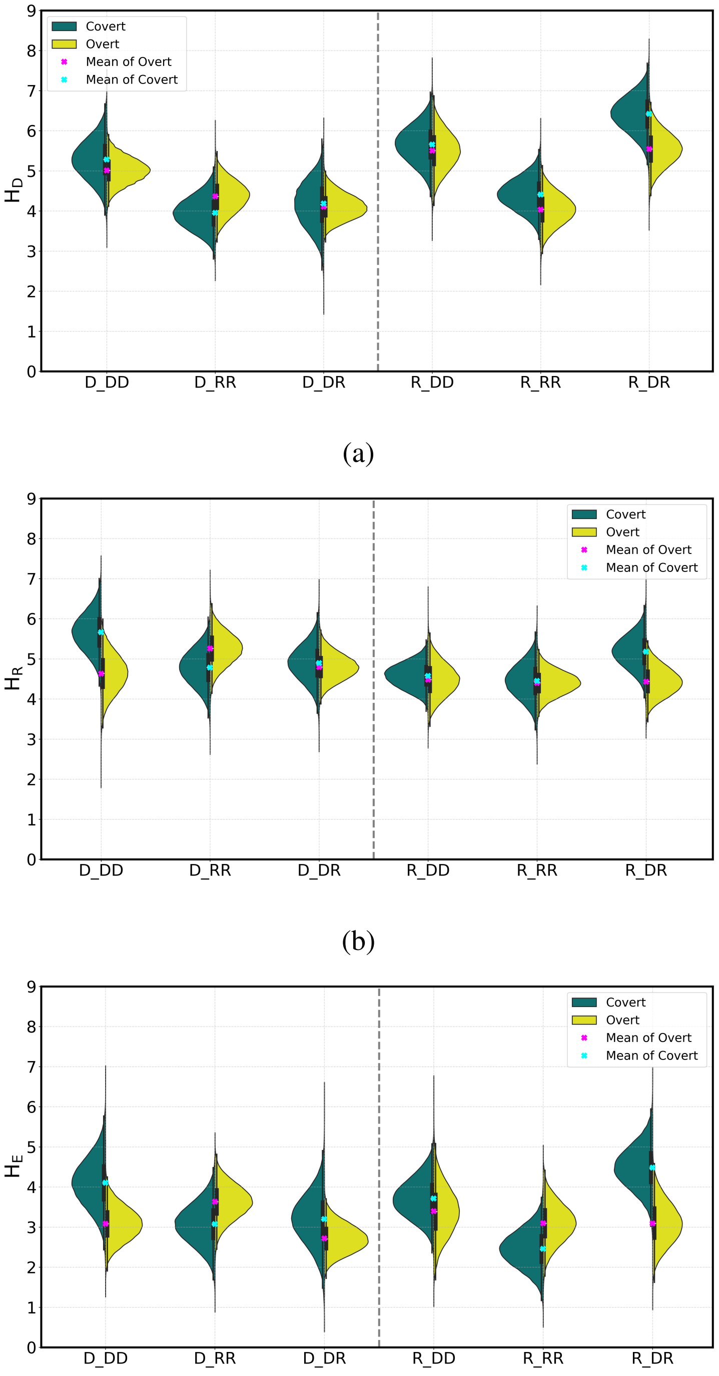


(a)

(b)

(c)

Figure B.4: Distributions of helpfulness scores from the bootstrapping process in notes written by team configuration and tweet source. Categories are distinguished by the partisanship of the tweet (D: Democrat, R: Republican) and the partisanship of the team members (DD: two Democrats, RR: two Republicans, DR: one Democrat and one Republican). The mean of the original data points is marked by the green x. **Panel (a):** The helpfulness scores as evaluated by self-identified Democrats. **Panel (b):** The helpfulness scores as evaluated by self-identified Republicans. **Panel (c):** The helpfulness scores as evaluated by experts.



(a)

(b)

(c)

Figure B.5: Comparison of the distributions of helpfulness scores from the bootstrapping process in notes written by teams v. individuals, categorized by team configuration and tweet source. Categories are distinguished by the partisanship of the tweet (D: Democrat, R: Republican) and the partisanship of the team members (DD: two Democrats, RR: two Republicans, DR: one Democrat and one Republican). The teams in the covert and overt treatments are shown separately. The mean of the original data points is marked by the green x. **Panel (a):** The helpfulness scores as evaluated by self-identified Democrats. **Panel (b):** The helpfulness scores as evaluated by self-identified Republicans. **Panel (c):** The helpfulness scores as evaluated by experts.

Appendix C

Supplementary Material for Chapter 4

C.1 Pre-game questionnaires

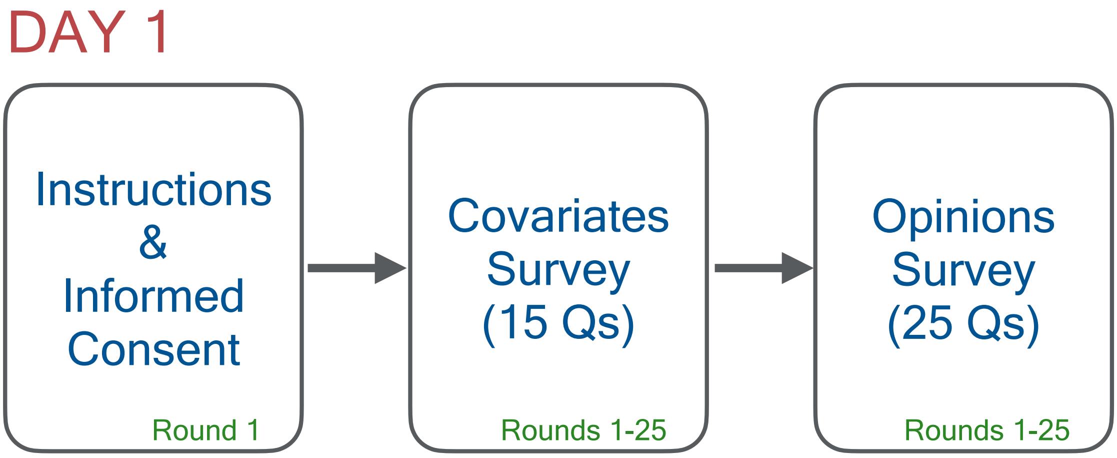


Figure C.1: Experimental flow for day 1

C.1.1 Instructions and informed consent for day 1

Welcome to the first phase of EXP12-Sapphire!

Thank you for participating in this study! This experiment has been conducted by Gabriela Juncosa as part of her doctoral thesis project, which explores the dynamics of creating connections. You must be at least 18 years old to participate in this study. Participation in this study more than once is not allowed.

This experiment will take place over two consecutive days. On the first day, you will be asked to provide your ideas and opinions on various topics of interest through a series of questions.

Today's session will take approximately 20-30 minutes. On the second day, you will have the opportunity to participate and compare your views with those of other study participants. The second session will take place tomorrow, [DATE], at [TIME] (Madrid time) and will last approximately 1 hour.

The payment for completing the first part will be €2; however, to receive payment, you must complete both parts to the end. Payment for all participants will be made through PayPal in the week following the completion of the survey. If you do not have a PayPal account, we will not be able to pay you.

You must participate in the experiment from a COMPUTER. The interface will not work on other devices such as tablets or phones.

If you wish to withdraw at any time, you can simply close the browser. If you wish to withdraw your data from the study or encounter any issues with the experiment, please contact juncosa_maria@phd.ceu.edu or ibsen.gisc@gmail.com.

By clicking "Next," you acknowledge that you have read and understood the above and wish to participate in this study.

C.2 Covariates survey

Table C.1: **Day 1 personality traits question wordings.** Questions 1–14 were measured on a 5-point Likert scale, with 1 indicating strong disagreement and 5 indicating strong agreement. Question 15 required a binary Yes/No response. Where multiple statements measured the same variable, responses were averaged to create a single index. (*) indicates that the scale was reversed to maintain consistency across statements.

Variable		Question
Predisposition to Share	1	Imagine you are at a party where you don't know most of the people. You are talking to a group of people when someone brings up the topic of same-sex marriage. From the discussion you can deduce that most of the people in the group do not support your point of view. In this type of situation, some people would express their opinion and others would not. How likely are you to express your opinion in such a situation?
Fear of Isolation	2	I worry about being isolated if people don't agree with me
	3*	I don't worry about people avoiding me
Conflict avoidant	4	I avoid telling others what I think when there is a risk that they will avoid me if they know my opinion
	5	Discussing controversial topics improves my intelligence
	6*	I enjoy a good discussion on a controversial topic
	7	I try not to get into arguments
Need to Evaluate	8	I like to have strong opinions even when I am not personally involved
	9	I have a judgment and an opinion about everything
	10	For me, it is very important to have strong and firm opinions
	11	It bothers me to be neutral
	12	I have many more opinions than the average person
	13	I prefer to have a strong opinion rather than no opinion at all
Engaged in Politics	14	I keep up to date with news about politics, public policy, or controversial issues such as race, gender, or immigration
Active in Social Media	15	During the past 12 months, have you posted public content (message, post, comment) on social media (Facebook, X, YouTube, Instagram, Snapchat, TikTok, etc.) about politics, public policy, or controversial social issues such as race, gender, or immigration?

Table C.2: **Day 2 outcome variables question wordings.** On Day 2, participants were first given the opportunity to change their answer to the opinion statements from the previous day. Next, they were explicitly asked whether they wanted to share their opinions with their connections. Both questions had binary Yes/No response options.

Variable	Question
Opinion change	Yesterday, your reaction to the above statement was: [AGREE/DISAGREE]. Would you like to change your answer?
Willingness to share	Based on your answer above, the following statement will be shared with your connections: [INSERT STATEMENT]. Would you like to share this opinion with your contacts?

C.3 Opinion survey

Table C.3: **Topic-specific willingness-to-share covariates question wordings.** On Day 1, participants answered questions measuring their attitude certainty, issue importance, and perception of the majority climate. These questions were presented before the interactive portion of the experiment to minimize priming and bias. On Day 2, participants were asked whether they believed the issue in question was important to their connections. All answers were measured on a 5-point Likert scale, with 1 indicating strong disagreement and 5 indicating strong agreement.

Day	Variable	Question
1	Attitude Certainty	I have a strong conviction regarding my position on this issue
	Issue Importance to Ego	This topic is very important to me personally
	Majority Climate	My opinion on the subject is similar to most of the opinions I hear in my
2	Issue Importance to Connections	This topic is very important to my connections

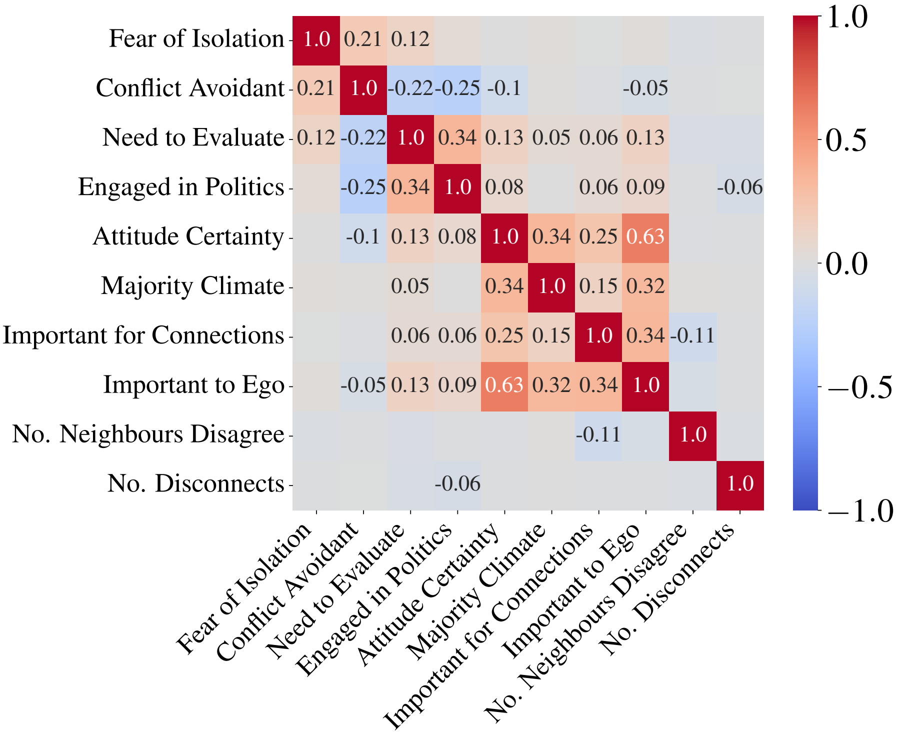
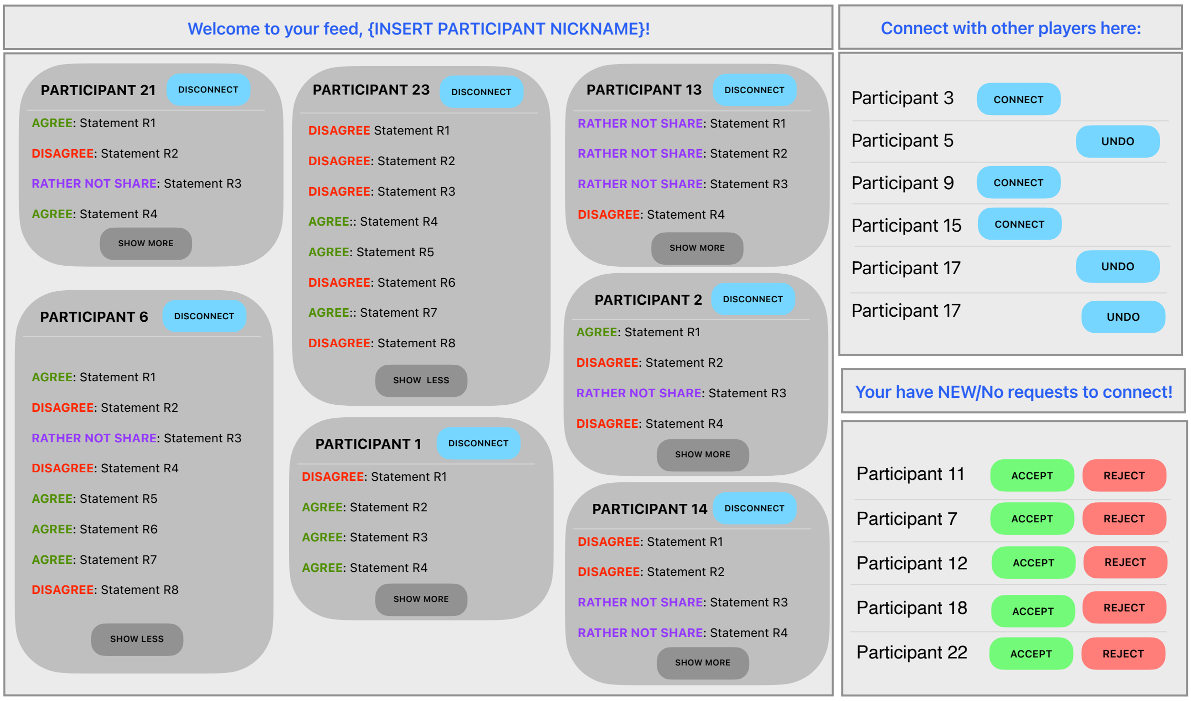


Figure C.2: **Correlation Matrix of Predictor Variables.** The figure displays pairwise correlations among key predictor variables. Correlation coefficients range from -1 (strong negative correlation) to 1 (strong positive correlation), represented by both numerical values and color intensity. Warmer colors (red) indicate positive correlations, while cooler colors (blue) represent negative correlations. Low correlation coefficients (between -0.05 and 0.05) are omitted for clarity.

C.4 Connections game

Table C.4: Number of participants per session recruited and those who completed the session . Dropout rates for interactive portion of the experiment (Day 2).

Session	Recruited	Completed	%
1	29	26	10.3
2	30	26	13.3
3	32	26	18.8
4	29	22	24.1
5	30	24	20.0
6	28	26	7.1
7	32	27	15.6
8	29	23	20.7
9	29	23	20.7
10	30	29	3.3



Welcome to your feed, {INSERT PARTICIPANT NICKNAME}!

PARTICIPANT 21 DISCONNECT
AGREE: Statement R1
DISAGREE: Statement R2
RATHER NOT SHARE: Statement R3
AGREE: Statement R4
SHOW MORE

PARTICIPANT 6 DISCONNECT
AGREE: Statement R1
DISAGREE: Statement R2
RATHER NOT SHARE: Statement R3
DISAGREE: Statement R4
AGREE: Statement R5
AGREE: Statement R6
AGREE: Statement R7
DISAGREE: Statement R8
SHOW LESS

PARTICIPANT 23 DISCONNECT
DISAGREE Statement R1
DISAGREE: Statement R2
DISAGREE: Statement R3
AGREE:: Statement R4
AGREE: Statement R5
DISAGREE: Statement R6
AGREE:: Statement R7
DISAGREE: Statement R8
SHOW LESS

PARTICIPANT 1 DISCONNECT
DISAGREE: Statement R1
AGREE: Statement R2
AGREE: Statement R3
AGREE: Statement R4
SHOW MORE

PARTICIPANT 13 DISCONNECT
RATHER NOT SHARE: Statement R1
RATHER NOT SHARE: Statement R2
RATHER NOT SHARE: Statement R3
DISAGREE: Statement R4
SHOW MORE

PARTICIPANT 2 DISCONNECT
AGREE: Statement R1
DISAGREE: Statement R2
RATHER NOT SHARE: Statement R3
DISAGREE: Statement R4
SHOW MORE

PARTICIPANT 14 DISCONNECT
DISAGREE: Statement R1
DISAGREE: Statement R2
RATHER NOT SHARE: Statement R3
RATHER NOT SHARE: Statement R4
SHOW MORE

Connect with other players here:

Participant 3	CONNECT
Participant 5	UNDO
Participant 9	CONNECT
Participant 15	CONNECT
Participant 17	UNDO
Participant 17	UNDO

Your have NEW/No requests to connect!

Participant 11	ACCEPT	REJECT
Participant 7	ACCEPT	REJECT
Participant 12	ACCEPT	REJECT
Participant 18	ACCEPT	REJECT
Participant 22	ACCEPT	REJECT

Figure C.3: Illustration of connection game's user interface (day 2)

C.5 Instructions and informed consent for day 2

Welcome to the first phase of EXP12-Sapphire!

Thank you for participating in this study! This experiment has been conducted by Gabriela Juncosa as part of her doctoral thesis project, which studies social dynamics. You must be at least 18 years old to participate in this study. Participation in this study more than once is not allowed.

In today's session, you will have the opportunity to interact with other participants and compare your views with others in the study. To see others' opinions, you need to first connect with them. The experiment begins with some randomly created connections, and from there, you can create new connections with any participant. To create a connection, you must first send a request (click on ”CONNECT”) and wait for the other participant to accept (they will need to click ”ACCEPT”) or reject (they will click ”REJECT”) the request. You can also choose to cancel a request (click ”UNDO REQUEST”). You may also receive connection requests from other participants and accept or reject them as explained earlier. If the request is accepted, both you and your partner can see each other's opinions in the current round and any previous rounds in which they were connected. At any time, either of you can disconnect a connection by clicking ”DISCONNECT”.

The messages in your connection histories reflect the agreement or disagreement between your opinions and those of your partners. For example, if you and your partner both agreed with a statement, the message will be ”Agree with you.” If both you and your partner disagreed with the statement, the message will also be ”Agree with you.” The message ”Disagree with you” can mean that you indicated agreement and your partner disagreed, or vice versa. In the table below, you have all the possible messages and their meanings. It is recommended to take a screenshot of this table now.

If you don’t remember the statement to which the message refers, hover your cursor over the corresponding round, and a window with the associated statement will appear, as shown in the image below.

If you don’t remember your opinions, the left panel contains a history of your responses. This panel is visible only to you.

Payment will be proportional to the points you accumulate throughout the game. For each point you earn, you will receive €0.04, and we estimate the average payment for all participants to be €15. There are two ways to earn points. First, you will receive 1 point for each connection you have at the end of a round. Second, you will receive extra points for accurately assessing whether your opinion is in the minority or majority compared to your connections. We have illustrated how to use the information displayed to evaluate the opinions of your connections. Please make sure you have taken a screenshot of the table above.

The experiment consists of 25 rounds, and in total, the study will last about 60 minutes. Each screen in the experiment has a time limit. If you do not act within the allotted time, we will randomly select a response/action for you. If you remain inactive for 3 rounds or more, you will not be able to continue with the experiment and will not receive any payment.

You must participate in the experiment ONLY from a COMPUTER. The interface will not work on other devices such as tablets or phones.

If you wish to withdraw at any time, you can simply close the experiment window. In that case, you will not receive any payment. If you wish to withdraw your data from the study or encounter any issues with the experiment, please contact juncosa_maria@phd.ceu.edu or ibsen.gisc@gmail.com. By clicking ”Next,” you acknowledge that you have read and understood the above and wish to participate in this study.”

C.6 Experiment dynamics and survey responses statistics

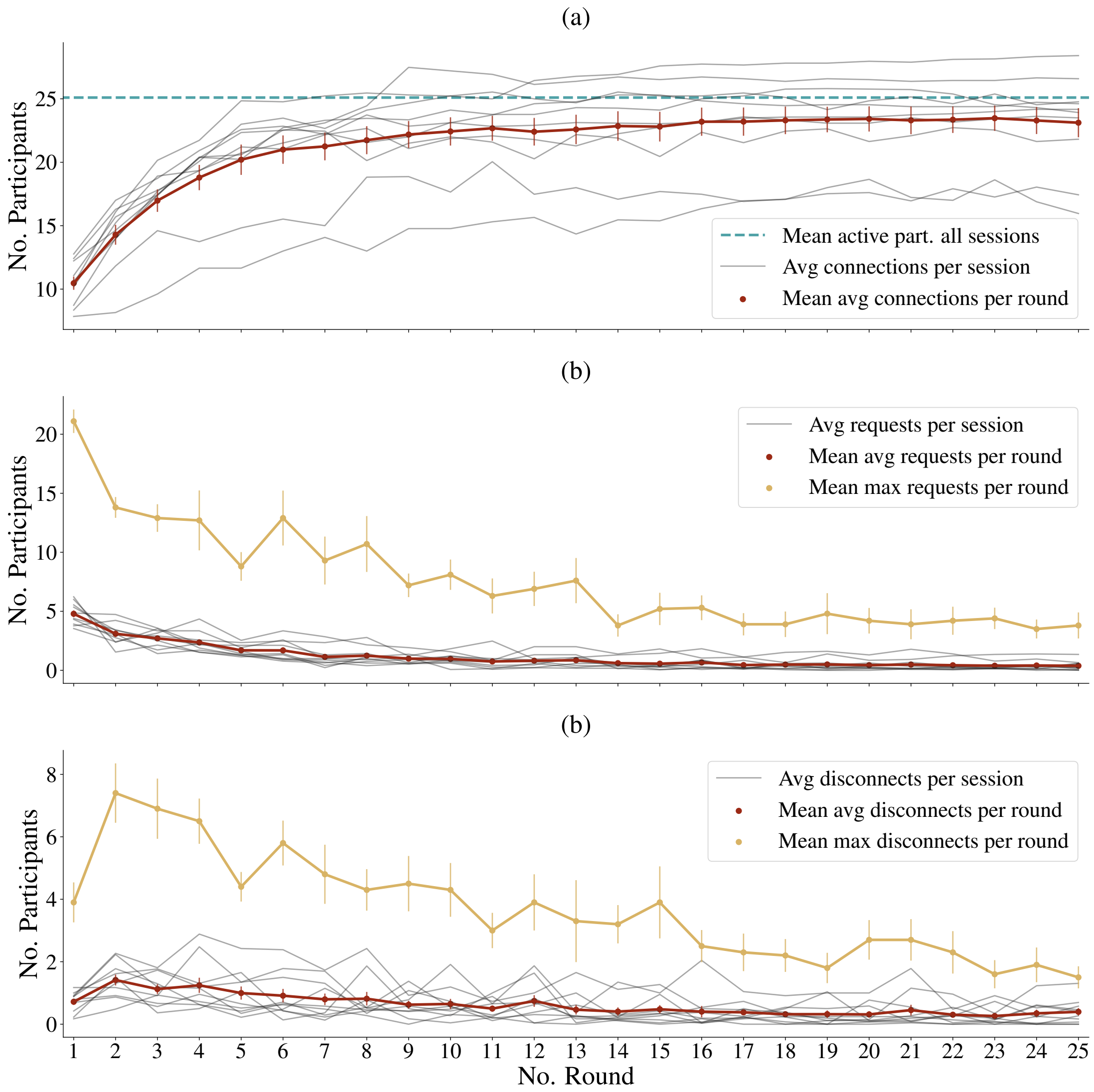


Figure C.4: Overview of connections game dynamics. Panel (a) shows the number of connections per participant averaged over all participants in that session (gray lines). Additionally, it shows the mean average connections for all rounds (red line). As expected, mean average connections converges quickly to the mean active participants (blue dotted line), suggesting that participants strive to connect to others. Panel (b) shows statistics for the number of requests sent. Here, we observe that requests are mostly sent at the beginning of the session. Panel (c) Shows statistics for the number of disconnects, although still low, this behavior exhibits more heterogeneity that request sending.

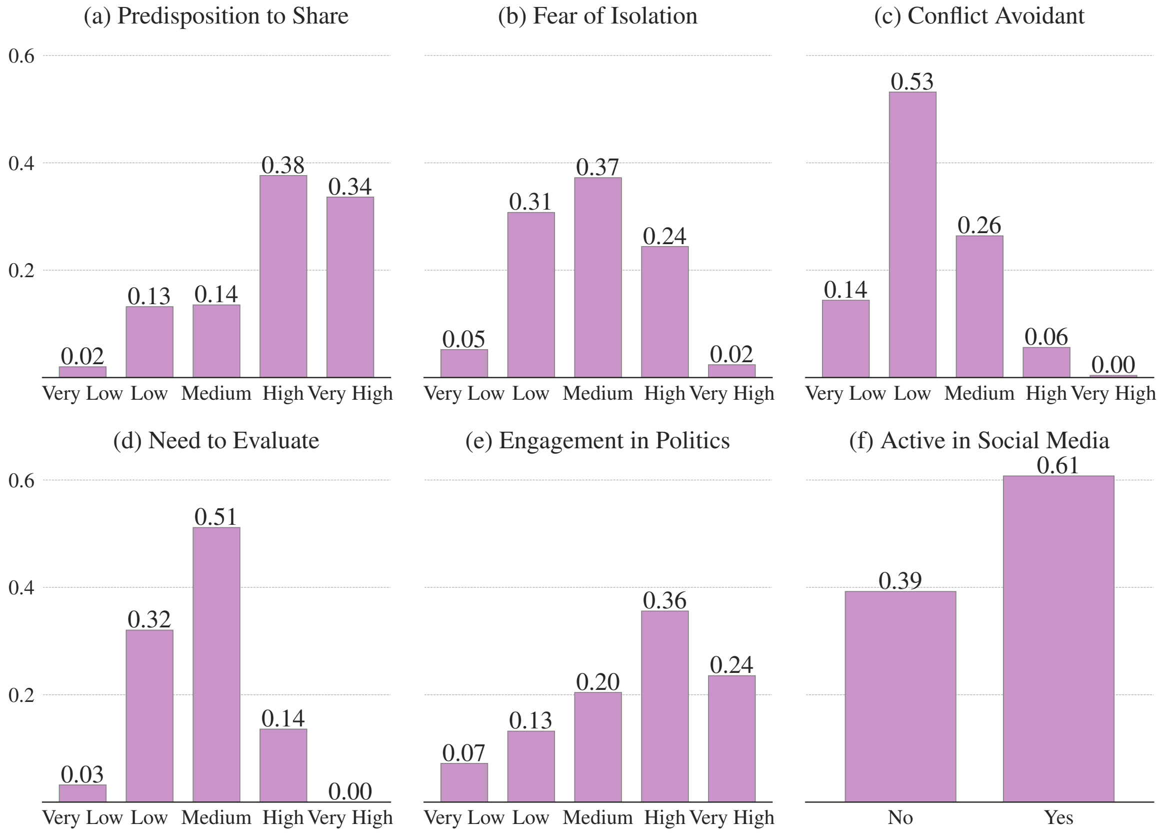


Figure C.5: Descriptive statistics for Day 1 personality traits survey

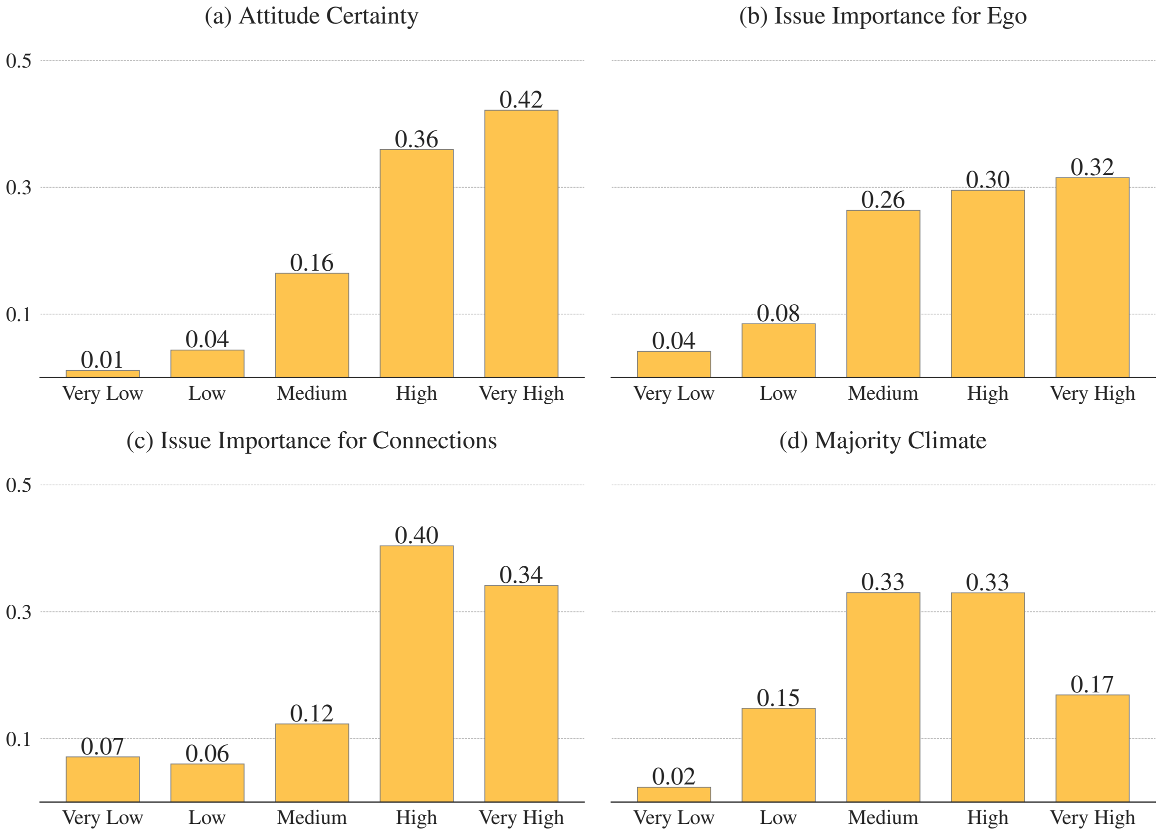


Figure C.6: Descriptive statistics for Spiral of Silence covariates survey

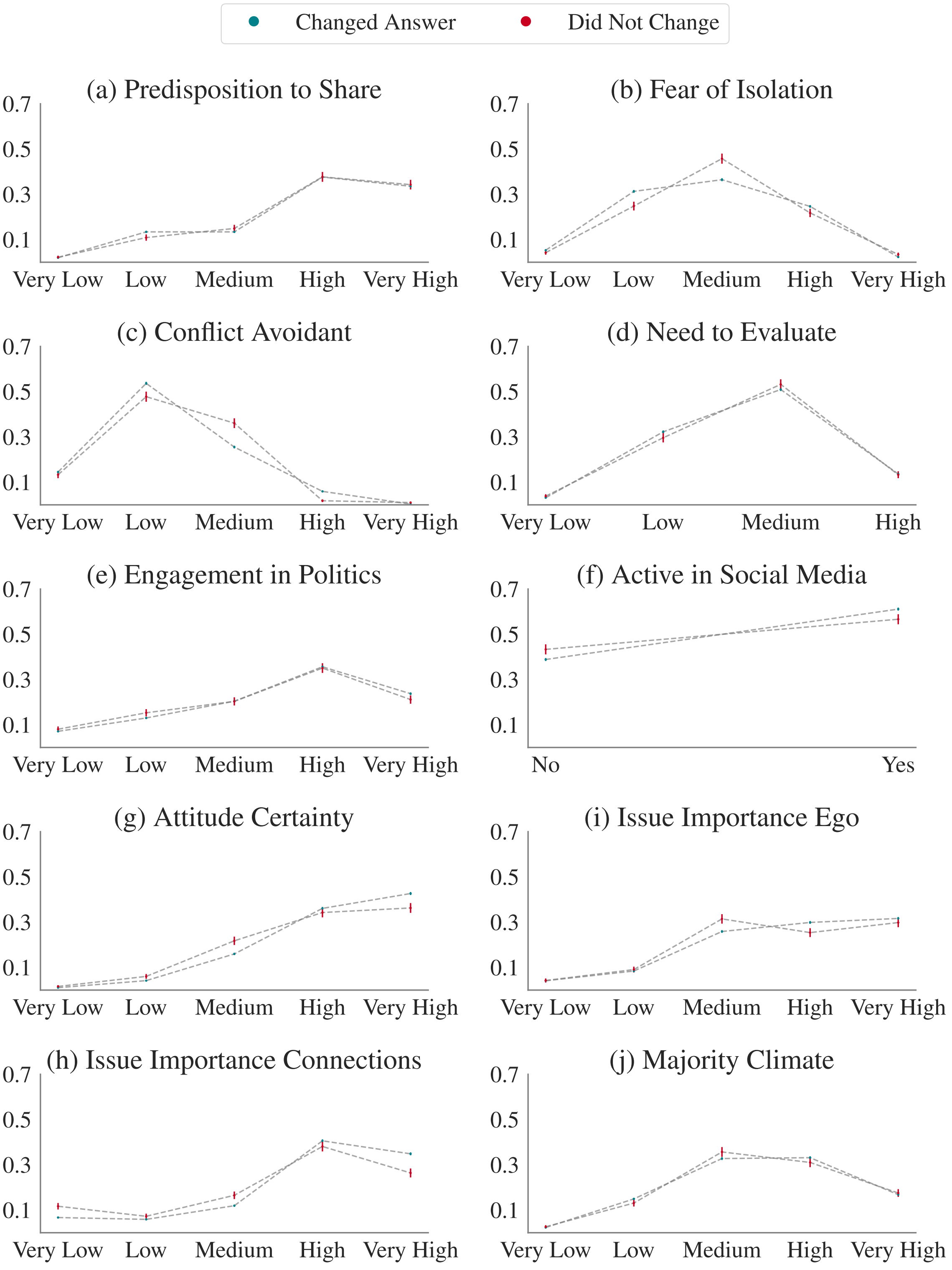


Figure C.7: Correlation between decision to change opinion and Spiral of Silence covariates

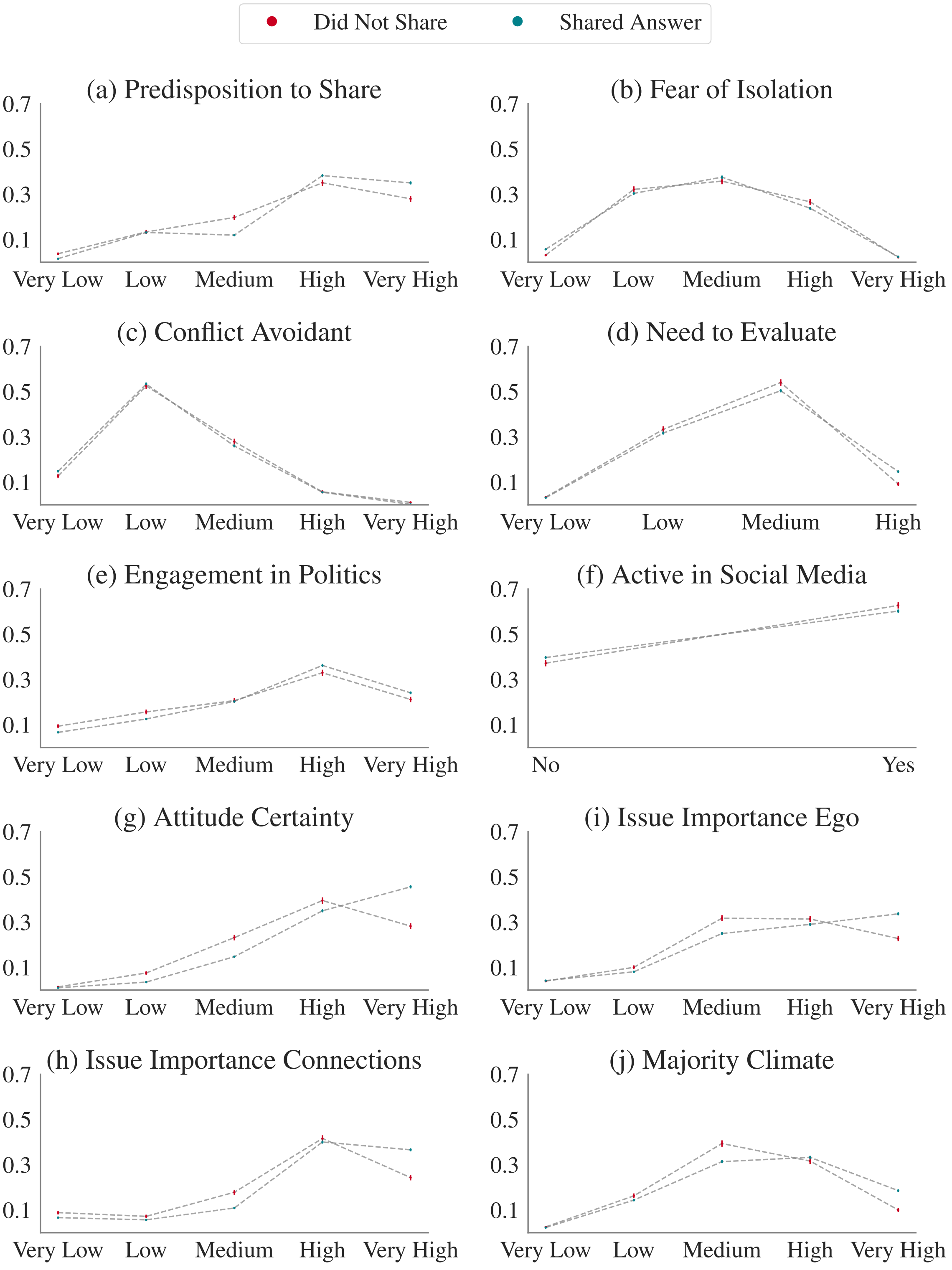


Figure C.8: Correlation between willingness to share and Spiral of Silence covariates