UNIVERSITY OF CALIFORNIA, MERCED

Characterizing Language & Users on Social Media

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Applied Mathematics

by

Maia M. Powell

Committee in charge:

Prof. Arnold Kim, Chair

Prof. Suzanne Sindi

Prof. Erica Rutter

2024

The dissertation of Maia M. Powell is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

(Prof. Suzanne Sindi)

(Prof. Erica Rutter)

(Prof. Arnold Kim, Chair)

University of California, Merced

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

To my fellow graduate students, this Ph.D. is as much yours as it is mine. Thank you for the cram sessions on holidays, the unflinching help with coursework at unconventional hours, and the endless hours of Zoom venting. For joining me in doing math—sometimes while laughing, sometimes while crying, and often times both—thank you. My journey would not have been difficult without you, it would have been impossible.

To my committee, thank you for your limitless patience, steadfast belief in me, and for always being in my corner. I will carry your teachings with me for the rest of my life, through every challenge and triumph, both personally and professionally.

To my friends and family, I love you more than words can describe. Thank you for bringing light into my life and for always celebrating me.

Maia Powell

Applied Mathematics Ph.D.
University of California, Merced

☎ *719-930-9059*
✉ *mpowell2@ucmerced.edu*
`www.mmpowell.com`

Research Interests

data science, natural language processing, social media, sociology, social good/justice

Education

Summer 2024 **University of California, Merced**, *Doctor of Philosophy*, Applied Mathematics
Dissertation Title: "Characterizing Language & Users on Social Media"
Advisor: Dr. Arnold Kim

Spring 2018 **University of Northern Colorado**, *Bachelor of Science*, Applied Mathematical Sciences
Concentration: Computer Science; Minor: Sociology

Work Experience

Summer 2023 **Data Science for Social Good Fellow**, *eScience Institute - University of Washington*
Computed groundwater anomalies, developed open-source workflow, collaborated with interdisciplinary team, collected and integrated satellite and climate/weather data using APIs, created geospatial visualizations, communicated findings to stakeholders to inform sustainable groundwater policy

Summer 2022 **Computing Intern**, *Lawrence Livermore National Lab*
Researched and implemented information extraction techniques for PDF documents, created and optimized/enhanced extraction pipelines, effectively communicated findings to leadership

Winter 2021 **Math-to-Industry Boot Camp Participant**, *Institute for Mathematics and its Applications, University of Minnesota*
Forecasted corn yield, integrated climate data and agricultural data through API scraping, produced dynamic geospatial visualizations, collaborated virtually with interdisciplinary team

Summer 2021 **Data Science Summer Institute Graduate Intern**, *Lawrence Livermore National Lab*
Created social media datasets for analysis, applied machine learning methods to analyze and classify social media data, utilized computing resources
Natural Language Processing Reading Group Cohort Lead - prepared weekly materials, led discussions

Summer 2020 **Data Science Challenge Team Lead**, *Lawrence Livermore National Lab*
Led and managed multidisciplinary team of undergraduate students, delegated coding work, applied machine learning classification methods, researched chemical properties

Summer 2019 **Graduate Student Researcher**, *University of California, Merced*
Studied and implemented natural language processing methods with emphasis on social media data, explored word/sentence embedding techniques, scraped and processed data from the Twitter API, analyzed text/corpus, user, and network data, created and deployed surveys, analyzed survey data

Fall 2018 - Summer 2019 **Teaching Assistant**, *University of California, Merced*
Courses: Calculus I, in support of Professor Alexander Yatskar, Linear Algebra & Differential Equations, in support of Professor Li-Hsuan Huang
Facilitated classroom activities and discussions, graded student assignments and exams, maintained accurate records, provided feedback to students, delivered course review lectures, held office hours

Honors & Awards

Fall 2023 - Spring 2024 **Research Training Grant Fellowship**, *National Science Foundation*, University of California, Merced

Spring 2023 **2023 Mathematically Gifted and Black Rising Star**, *Mathematically Gifted and Black*, Network for Minorities in Mathematical Sciences

Spring 2022 **GradSlam Finalist**, *University of California*, University of California, Merced

Spring 2020 - Summer 2023 **2020 Graduate Research Fellowship**, *National Science Foundation*, University of California, Merced

Fall 2018- Spring 2020 **National Research Traineeship - Intelligent Adaptive Systems**, *National Science Foundation*, University of California, Merced

Summer 2018 **Competitive Edge Summer Bridge Program Participant**, *UC Merced Graduate Division*, University of California, Merced

Service & Outreach

Spring 2023 **Graduate Dean's Advisory Council on Diversity Member**, *Graduate Division*, University of California, Merced

Collaborated with Graduate Dean and fellow graduate students to address issues of diversity, equity, and inclusion within the graduate student community; engaged in constructive dialogue

Fall 2022 - Spring 2023 **TEDx Conference Co-Organizer**, *TED*

Worked collaboratively, planned budgets, marketed event, coordinated volunteers, organized committees, curated content with emphasis on Central Valley research, managed tickets

Spring 2022 **Social Media & Graphic Design Specialist**, *Graduate Student Association*, Graduate Division, University of California, Merced

Maintained social media accounts, assisted collaborators with necessary graphic design needs

Fall 2021 - Spring 2024 **Peer Mentor**, *GradEXCEL Program*, Graduate Division, University of California, Merced

Provided guidance to first-year Ph.D. students, provided individualized support, organized and held regular meetings

Fall 2020 - Spring 2024 **Secretary, Vice President**, *RadioBio Podcast*, University of California, Merced

Interviewed invited guests, edited audio content, created art to improve science communication, recorded/maintained meeting notes, oversaw internal communications

Fall 2018 - Spring 2021 **Graduate Secretary, Graduate Vice President, Mentor**, *Women in STEM Program*, University of California, Merced

Maintained online presence of organization, advertised events, recorded/maintained database of meeting notes, coordinated internal program affairs, provided individualized guidance/support to mentee

Publications

Powell, Maia, Arnold D. Kim, and Paul E. Smaldino., "Hashtags as signals of political identity:#BlackLivesMatter and #AllLivesMatter." *Plos one* 18, no. 6 (2023): e0286524.

Padilla, Lace MK, **Maia Powell**, Matthew Kay, and Jessica Hullman., "Factors predicting willingness to share COVID-19 misinformation." *Frontiers in Psychology* 11 (2021): 579267.

Lobato, Emilio JC, **Maia Powell**, Lace MK Padilla, and Colin Holbrook., "Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations." *Frontiers in Psychology* 11 (2020): 566108.

Presentations

(Oral) "Flowing Forward: A Reproducible Workflow for Studying Groundwater Scarcity in the Colorado River Basin" presented at the Learning and Doing Data for Good Special Session; Academic Data Science Alliance Annual Meeting; October 25-27, 2023; San Antonio, Texas.

(Poster) "Evaluating Differences Between #BlackLivesMatter and #AllLivesMatter: Discourse and Interpretations" presented at Society for Industrial and Applied Mathematics (SIAM) Conference on Computational Science & Engineering; February 26-March 3, 2023; Amsterdam, the Netherlands.

(Invited, Oral) "Hashtags as Signals of Political Identity: the Case of #AllLivesMatter vs. #BlackLivesMatter" presented at the Workshop Celebrating Diversity: Data In Action; Society for Industrial and Applied Mathematics (SIAM) Annual Meeting; July 11-15, 2022; Pittsburgh, Pennsylvania.

(Poster) "Discourse analysis of pairwise Twitter hashtags" presented at Society for Advancement of Chicanos/Hispanics & Native Americans in Science (SACNAS) National Diversity in STEM Virtual Conference; October 22-24, 2020; Virtual.

Skills

Computing	Python (Jupyter, Pandas, Numpy, NLTK, Tweepy, Scitkit-learn, Scipy, Matplotlib, Seaborn)
	MATLAB, R
	Bash scripting, Git/Github
	LaTeX, Microsoft Office Suite

ABSTRACT OF THE DISSERTATION

Characterizing Language & Users on Social Media

by Maia M. Powell

Doctor of Philosophy in Applied Mathematics

University of California Merced, 2024

Committee Chair: Prof. Arnold Kim

We explore mechanisms behind social media by leveraging diverse datasets to examine the intersection between users and language. Our exploration centers on understanding how demographics influence online behaviors, particularly in the context of high-engagement socio-political discourse (such as conversations surrounding #BlackLivesMatter) and the dissemination of misinformation related to COVID-19 on micro-blogging platforms. We aim to contribute insights for actionable strategies for promoting informed dialogue and combating misinformation, understand the complexities of online discourse, and shed light on the ways in which social identity shapes individuals' perceptions and participation in the digital realm.

Chapter 1

Overview

An overwhelming amount of conversations occur online. In fact, it is believed that most conversations take place online. The digital landscape has thus emerged as a dominant arena for human interaction, with a vast repository of data ripe for analysis (Kwak et al., 2010; Java et al., 2007; Bakshy et al., 2011). Online conversations provide invaluable insights into the dynamics of social interaction, offering researchers a window into the intricacies of human behavior in virtual spaces.

Our demographics not only make us unique, but affect our behaviors in surprising ways. The influence of demographics on our perceptions and behaviors is profound, extending beyond thoughts and opinions to shaping our active interactions with the world. Factors such as age, ethnicity, and socioeconomic status possess a surprising amount of influence over our behaviors. Studies have linked differential behaviors to differences in political affiliations (Young et al., 2019), consumer habits (Burkolter & Kluge, 2011), sustainability actions (Ajibade & Boateng, 2021), and even social media usage (Pan et al., 2017).

We are constantly signaling who we are, both in online forums and in real life, both intentionally and unintentionally. Whether through our attire or our linguistic style, we continuously broadcast signals that reflect our individuality and social affiliations. It is widely acknowledged by social scientists that an important function of public communication is to signal one's real or potential membership in some categorizable subset of individuals (Goffman, 1978; Loury, 1994; Donath, 1999; Berger & Heath, 2008; Smaldino, 2019). Identity signaling serves numerous social functions, such as indicating one's commitment to particular groups (Frank, 1988; Iannaccone, 1992; Sosis & Alcorta, 2003) and facilitating cooperative assortment for activities requiring cooperation or coordination (Smaldino, 2019; Barth, 1969; Nettle & Dunbar, 1997; McElreath, Boyd, & Richerson, 2003). Assortative signaling can be overt, so that information is widely received by diverse audiences, or covert, where information is encrypted so that only audiences "in the know" reliably perceive the identity-related content (Smaldino, Flamson, & McElreath, 2018; Smaldino

& Turner, 2021). Covert signals can be beneficial for the transmitter because they allow for individuals to strategically alter the clarity of their messages, imbuing them with cryptic or indirect meanings when they are likely to be viewed by hostile audiences (van der Does et al., 2022). Despite its use in facilitating cooperation between similar individuals, however, strategic identity signaling is not always aligned with societal good. For example, white supremacists have likely used covert signals on online social networks such as Twitter to coordinate with others while avoiding widespread detection (Bhat & Klein, 2020).

Unfortunately, not all content that is created online is positive. The internet has historically served as a hotbed for heated debate, sparking arguments and violent speech. In fact, individuals tend to be more willing to express controversial opinions online (Voggeser, Singh, & Göritz, 2018; Zimmerman & Ybarra, 2016). This is thought to be due to the inevitable distancing being online provides, coupled with the dehumanization of victims as a result (Whittaker & Kowalski, 2015). Further, anonymity afforded by online platforms emboldens hateful social media users to be more brazen and disseminate vitriolic rhetoric (Mondal et al., 2018).

Among the negative discourse spread on social media is misinformation, a phenomenon with far-reaching consequences. It has been found that misinformation spreads faster than true information (Del Vicario et al., 2016; Vosoughi, Roy, & Aral, 2018). Those spreading misinformation maliciously tend to draw and the novelty, shock, and emotion of misinformation to increase its potency and spread. Other sources suggest that some types of misinformation tends to appear lexically less complex than other types of information online (Carrasco-Farré, 2022; Charquero-Ballester et al., 2021), such as true information, making messages easier to digest, and thus spread, to lay audiences.

This thesis documents work I have done to understand the intersection of language and users on social media, seeking to identify the fundamental mechanisms driving online discourse. The work I have engaged in is characterized by a high degree of collaboration and interdisciplinarity, often intersecting with various branches of the social sciences. This collaborative approach has been instrumental in enriching the research process, as it brings together a diverse range of perspectives and expertise from multiple disciplines. The integration of these varied viewpoints has significantly enhanced the depth and quality of our findings.

I specialize in working with diverse types of data, including social media, network, text, and survey data. My work included not only data analysis, but additionally the mining, curation, and merging of such data for analysis. This includes the creation of surveys and querying of Application Programming Interfaces (APIs). By handling such a wide array of data types, I am able to develop a comprehensive and nuanced understanding of the online world. This multifaceted approach allows me to explore and interpret the complexities of digital interactions and behaviors from multiple angles, ultimately leading to more robust and insightful conclusions.

In summary, my research projects collectively explore the overarching question:

How does social identity shape individuals' perceptions and participation

online?

The remainder of this thesis is organized as follows: In Chapter 2, I will report work done on understanding misinformation spread on social media, specifically highlighting how demographics influence spread and how networks created from misinformation are characterized. In Chapter 3, I will report work done study viral socio-political movements. This work includes comparing differing discourse on microblogging platforms, examining how the use of hashtags can signal political identity, and how the volume of some discourse can affect the volume of another. Finally, in Chapter 4, I will state concluding remarks and present avenues for future, related research.

Chapter 2

Misinformation Surrounding COVID-19

2.1 Introduction

The COVID-19 pandemic, originating from the novel coronavirus SARS-CoV-2, has been one of the most significant global health crises in recent history. Beginning in late 2019 in the city of Wuhan, China, the virus quickly spread across borders, transcending geographical boundaries and impacting virtually every corner of the globe. The rapid transmission of the virus led to an unprecedented international response, with governments implementing various measures to curb its spread, including lockdowns, social distancing protocols, mask mandates, and travel restrictions (Ciotti et al., 2020; Shi et al., 2020). The pandemic brought about immense challenges, both in terms of public health and socio-economic impact. Furthermore, the economic ramifications of the pandemic were profound, with businesses shuttering, millions losing their jobs, and global supply chains disrupted (Béland, Brodeur, & Wright, 2023; Akbulaev, Mammadov, & Aliyev, 2020). Vaccination campaigns were launched worldwide, aiming to achieve herd immunity and bring an end to the pandemic; however, challenges such as vaccine hesitancy, inequitable distribution, and the emergence of new variants continued to pose obstacles to achieving this goal (Ki, 2021; Anderson et al., 2020; Dong, He, & Deng, 2021). The COVID-19 pandemic highlighted the critical role of science, healthcare infrastructure, and public health measures in mitigating the impact of infectious diseases. As communities continue to navigate the uncertainties posed by the pandemic, the lessons learned from this crisis will undoubtedly shape future preparedness efforts and response strategies to safeguard public health on a global scale.

The virus's impact was disproportionately severe on marginalized communities, including racial and ethnic minorities, low-income families, and those with pre-existing health conditions (Gauthier et al., 2021; Navarro & Hernandez, 2022; Bowleg, 2020; Strassle et al., 2022), starkly exacerbating existing disparities across various dimensions such as income, healthcare access, education, and employment. Structural inequities meant that these groups were more likely to be frontline workers, less able to work from home, and often resided in densely populated areas with limited access to healthcare. Consequently, they faced higher rates of infection and mortality, underscoring the critical intersections between health, socio-economic status, and racial inequality (Guasti, 2020; Reid, Ronda-Perez, & Schenker, 2021). The shift to remote work, while a viable solution for many white-collar professionals, was not feasible for numerous workers in service industries, manufacturing, and other sectors requiring physical presence. This dichotomy led to significant economic hardships for those unable to transition to remote work, resulting in job losses, reduced incomes, and financial instability (Gaitens et al., 2021). The pandemic thus exposed and intensified the economic vulnerabilities of already disadvantaged populations, highlighting the urgent need for systemic changes to address these deep-rooted inequities.

In parallel, the pandemic also catalyzed a disturbing rise in anti-Asian hate and xenophobia. With the origin of the virus traced to Wuhan, China, Asian communities worldwide became scapegoats, facing increased incidents of verbal abuse, physical assaults, and discrimination (Lantz & Wenger, 2023; Tessler, Choi, & Kao, 2020). Misguided rhetoric from political leaders and misinformation on social media platforms further fueled xenophobic sentiments, leading to a surge in hate crimes against Asian individuals (J. Y. Kim & Kesari, 2021). This wave of hostility not only endangered the physical safety and mental well-being of Asian communities but also reflected long-standing racial prejudices and the scapegoating of minority groups during times of crisis (Gover, Harper, & Langton, 2020). The pandemic, while devastating, offers a critical opportunity to confront and rectify the deep-seated inequalities and prejudices that continue to undermine social cohesion and justice.

The pandemic underscored the interconnectedness of the modern world (Mazzocchi, 2021). The utilization of social media platforms experienced a significant surge, attributed largely to the widespread imposition of lockdowns and social distancing measures. With traditional forms of social interaction and entertainment severely restricted, individuals increasingly turned to digital means of communication and engagement. Consequently, social media became a primary avenue for maintaining social connections, consuming news, and alleviating the pervasive sense of isolation, leading to unprecedented levels of online activity and user engagement (Aggarwal et al., 2022). Naturally, social media platforms provide fertile ground for the spread of misinformation due to their vast reach and the ease with which content can be shared. The phenomenon of misinformation and its defining characteristics have been particularly pronounced the context of the COVID-19 pandemic (Evanega et al., 2020; Cuan-Baltazar et al., 2020).

Social media platforms faced a significant challenge in moderating the widespread misinformation that proliferated online. Misinformation spreads on social media platforms like wildfire, often fueled by the rapid dissemination of information without proper fact-checking or vetting (Muhammed T & Mathew, 2022; Aımeur, Amri, & Brassard, 2023). One of the key drivers of misinformation about COVID-19 is the lack of centralized control over the content shared on social media platforms (Nsoesie et al., 2020). Platforms have taken steps to combat misinformation (Morrow et al., 2022), such as labeling disputed content, advanced algorithms to detect and flag false information, and implementation of stricter community guidelines to penalize repeat offenders. However, the sheer volume of information makes it difficult to moderate misinformation effectively. Moreover, the viral nature of social media means that false information can quickly gain traction and reach a wide audience before it can be debunked (Giansiracusa & Giansiracusa, 2021). This dynamic underscored the critical balance between maintaining open communication channels and ensuring public health and safety during a global crisis (Kozyreva et al., 2023).

2.2 Factors Predicting Willingness to Share COVID-19 Misinformation

Collaborators: Emilio Lobato, Prof. Lace Padilla, & Prof. Colin Holbrook

2.2.1 Motivation

There are numerous reasons why individuals might choose to spread misinformation (Balakrishnan et al., 2023). One significant factor is the exploitation of the uncertainty and fear surrounding the pandemic for various purposes. Some individuals and groups leverage this atmosphere of anxiety and confusion to advance political agendas, manipulating public opinion to gain power or discredit opponents (Ricard & Medeiros, 2020). Others see financial opportunities, such as selling fake cures or driving traffic to websites laden with advertisements, thereby profiting from the chaos (Pyzik et al., 2021; Papadogiannakis et al., 2023). Moreover, some people spread misinformation deliberately to create confusion and distrust in established institutions, aiming to undermine confidence in government agencies, health organizations (Diseases, 2020), and the media (Radwan, 2022). Conversely, there are those who share misinformation not out of malice but to seek verification or even to mock the absurdity of certain claims (Ahmed, Seguí, et al., 2020; Ahmed, Vidal-Alaball, et al., 2020; Brennan et al., 2020). These individuals may not intend to mislead but rather to question or highlight the ridiculous nature of some false information. However, even this unintentional spread can contribute to the overall dissemination of falsehoods. Additionally, a portion of the population genuinely believes in the misinformation they share. These individuals may be convinced of the accuracy of the false information and spread it with the intention of informing and protecting others. Their actions, while not malicious, still perpetuate false narratives and contribute to the problem. Regardless of the underlying motivation, the spread of misinformation is undeniably harmful and the ramifications of widespread misinformation are profound.

In the context of a pandemic, misinformation can lead to widespread public confusion and fear, causing people to disregard expert advice and health guidelines (Lăzăroiu, Mihăilă, & Branişte, 2021). This can result in lower vaccination rates (Pierri et al., 2022), the refusal of life-saving treatments, and the adoption of ineffective or dangerous remedies (Reihani et al., 2021). The proliferation of false information can additionally overwhelm healthcare systems. Many adopted alarming behaviors such as people drinking bleach and other harmful substances, partly fueled by then-President Donald Trump's comments suggesting that injecting disinfectants might treat the virus (Yamey & Gonsalves, 2020). Despite immediate clarifications from medical experts and health authorities debunking these claims, the misinformation spread rapidly across social media and other platforms. As a result, poison control centers reported spikes in cases of individuals ingesting bleach and other toxic substances, causing severe health complications and even fatalities (Kuehn, 2020; Chary et al., 2021). Beyond health impacts, misinformation can be weaponized to polarize communities, exacerbate divisions, and manipulate public opinion (Jiang, Ren, Ferrara, et al., 2021). This environment fosters a culture of misinformation where individuals become more isolated from differing perspectives and less capable of critical thinking. Combating misinformation requires concerted efforts from individuals, communities, and institutions to promote media literacy, encourage critical thinking, and foster environments where accurate information can thrive.

Misinformation differs from reliable information in more ways than just epistemic, allowing researchers to make finer-grained distinctions between types of misinformation. Research has found that conspiratorial misinformation differs along a number of dimensions from other categories of misinformation. Misinformation about COVID-19 takes various forms, ranging from conspiracy theories about the virus's origins to false claims about potential cures or prevention methods. The spread of COVID-19 conspiracies online is of particular importance because research has found a positive relationship between trusting information on social media and accepting COVID-19 conspiracies, as well as a negative relationship between trusting social media for COVID-19 information and health-protective behaviors (Allington et al., 2021; Bridgman et al., 2020; van Mulukom, 2020).

2.2.2 Approach

Primary objectives for this research are to examine how individual difference variables predict information sharing behaviors. The current work does not focus on the specific motivations people may have for sharing misinformation, but rather the overall willingness to share claims regarding the current COVID-19 pandemic that happens to be untrue or unverifiable over social media. The goal of the present exploratory research is to begin characterizing the socio-cognitive profile of individuals likely to spread misinformation online.

2.2.3 Data

We used fact-checking sites, such as Snopes.com and FactCheck.org, to create an *ad hoc* measure of peoples' willingness to share misinformation about COVID-19 over social media. Eighteen actual claims, either verified to be untrue or unverifiable, that have been made regarding COVID-19 were presented to participants. For each claim, participants used a slider to indicate how likely they would be to share that claim over their social media accounts. The slider bar ranged from scores of 0 to 100, with anchors of "Definitely not share," "Less likely to share," "More likely to share," and "Definitely share" located at the 0, 33, 66, and 100 marks, respectively. We calculated mean scores for participants' willingness to share misinformed claims about COVID-19. The items selected for this scale were *a priori* categorized as claims regarding: (a) severity and spread of COVID-19, (b) treatment and prevention of COVID-19, (c) COVID-19 conspiracy theories, and (d) miscellaneous incorrect or unverifiable claims. The categorization scheme utilized in the current work was based on the categorization structure of claims from the originating fact-checking sites [1].

2.2.4 Individual Differences

We explored whether different patterns of individual differences predict the inclination to share different kinds of misinformation about a salient socio-cultural scientific topic (the global COVID-19 pandemic). For the purposes of the present research, we limited our focus to individual differences in propensity toward conspiracy ideation, attitudes toward science, and facets of political ideology. Each of these individual differences has been previously found to relate either to the endorsement of misinformation or to how people respond to health threats from pathogens.

Conspiracy theorists typically posit explanations for large-scale events that contradict official or expert explanations (Goertzel, 1994). They tend to be distrustful of recognized legal or scientific cultural authorities. This distrust of authority is so pervasive in conspiracy ideation that people inclined to believe conspiracies will accept mutually exclusive conspiracy theories more than the official account of a major socio-cultural event (Wood, Douglas, & Sutton, 2012). Accordingly, we investigated the influence of individual differences in conspiracy ideation on willingness to share misinformation. We measured participants' disposition toward conspiracy ideation with the Conspiracy Mentality Questionnaire (Bruder et al., 2013). Participants rated their level of certainty about various statements on an 11-point Likert scale (0% – Certainly Not to 100% – Certain).

Researchers have found that belief in conspiracies correlates with the rejection of science and endorsement of pseudoscience (Lewandowsky, Gignac, & Oberauer, 2013; Lewandowsky, Oberauer, & Gignac, 2013; E. Lobato et al., 2014; Van der Linden, 2015; E. J. Lobato & Zimmerman, 2019) and to a general attitude toward science as lacking credibility. Therefore, understanding who is likely to spread misinformation about a scientific topic requires assessing attitudes about science in general. We measured participants' general attitudes toward science

[1] The categorization scheme in this study was inspired by categorizations used on Snopes.com: "Origins and Spread," "Treatment and Prevention," and "Conspiracy Theories." We build on this by including a "Miscellaneous" category which includes claims from diverse categories on the Snopes collection webpage, such as "Media and Entertainment" or "Prophecies and Predictions."

with the Credibility of Science Scale (CoSS) (Hartman et al., 2017). This six-item measure asks participants to respond on a 7-point Likert Scale (1 = Disagree Very Strongly; 7 = Agree Very Strongly), scored such that higher scores represent less favorable views of science as credible.

We include both measures of social dominance orientation and traditionalism to explore their relative contributions to the spread of health-related misinformation in the midst of a global pandemic, as the relationship between pathogen sensitivity and political views is driven primarily by ideologies favoring hierarchical social stratification. We used the Social Dominance Orientation short form (Pratto et al., 2013) to measure approval of social hierarchies. Participants respond to this four-item measure by using a 7-point Likert scale (1 = Extremely Oppose; 7 = Extremely Favor) to indicate how much they reject or support statements concerning social hierarchies and egalitarianism. We used the six-item Traditionalism subscale from the Authoritarian-Conservatism-Traditionalism scale (Duckitt et al., 2010) to measure participants' valuation of traditional moral systems and lifestyles and resistance to modern challenges to such traditional values and lifestyles. Participants responded on a 7-point Likert scale (1 = Strongly Disagree; 7 = Strongly Agree) to measure this.

We used a modified version of the Political Issues Index (Dodd et al., 2012; Holbrook, Lopez-Rodriguez, & Gomez, 2018) as a proxy for where participants generally fall on the liberal-to-conservative political spectrum. This 20-item measure lists socio-political issues (e.g., "Same-sex marriage," "Reduce business regulations," and "Right to abortion"), and participants indicate whether they Agree, Disagree, or are Uncertain about the issue. The Political Issues Index is scored from 1 to 1, reverse-scoring agreement with the traditionally liberal items, such that lower values represent greater alignment with traditionally liberal policy positions, and higher values represent greater alignment with traditionally conservative policy positions ("Uncertain" responses are scored as zero).

2.2.5 Survey

We recruited 404 participants via Amazon's Mechanical Turk. We removed data on the basis of preregistered criteria: incomplete responses to the dependent measure or individual difference measures, completing the study in less than 2 min, and failure to respond or nonsensical response to an open-ended question asking them to describe the study. The final sample, after exclusions, was 296 participants.

After providing informed consent, participants were presented with the following instructions:

> We are interested in examining what types of things people share over social media. Sometimes people share information because they think it is true and want others to know it. Sometimes people share information even if they think it is false because they would like to warn other people to not believe it if they hear it from somewhere else. Sometimes people share information that they are not sure about as a way to

> see what their friends and family think. And sometimes people share information for other reasons entirely.
>
> In this task, you will be presented with a series of claims regarding the current COVID-19 (aka SARS-CoV-2) pandemic that have been made and shared over both traditional media outlets, such as TV news programs or newspapers, and over social media outlets, such as Facebook or Twitter. You may have even encountered some of these already.
>
> For each claim, use the slider bar provided to rate how likely you think you would be to share this over your own social media accounts.

After reading the instructions, participants completed the task. The 18 claims we used as stimuli were presented in a randomized order. Participants were informed that these were real claims that have been made on both traditional news media outlets and on social media platforms. Following this task, participants filled out the individual difference measures in randomized order. Finally, participants filled out a demographics form. Participants were debriefed as to the nature of the study and informed that the claims they read regarding COVID-19 were not true. In the debriefing, we provided links to fact-checking and health agency websites for participants, to help provide participants with resources to keep up to date with COVID-19 information and misinformation.

2.2.6 Result & Discussion

We assessed the relationship between individual difference measures and self-reported willingness to share different kinds of COVID-19 misinformation over social media using a canonical correlation, allowing for analysis of the relationship between sets of predictor and outcome variables by creating synthetic variates representing linear combinations of the predictor variables and linear combinations of the outcome variables. This analysis strategy is designed to generate the highest correlation between the two variable sets (Tabachnick & Fidell, 2007).

The first model reveals that participants who are primarily more liberal (in terms of the issues index) and less oriented toward social dominance were less inclined to share COVID-19 claims that were conspiratorial in nature. The second model produced by the canonical analysis revealed that individuals high in Social Dominance Orientation and low in Traditionalism were less inclined to share misinformation claims regarding the severity and spread of COVID-19, but more inclined to share COVID-19 conspiracies and miscellaneous COVID-19 misinformation claims.

Overall, our canonical model revealed two distinct profiles predicting two patterns of willingness to share misinformation. The significant structure coefficients for both profiles hint that the relationships between the selected individual difference variables and the subtypes of COVID-19 misinformation studied here are more complicated than could be revealed by the use of a general linear model approach. Although every individual difference selected for inclusion in the present study was motivated by relevant prior literature, follow-up research is needed to

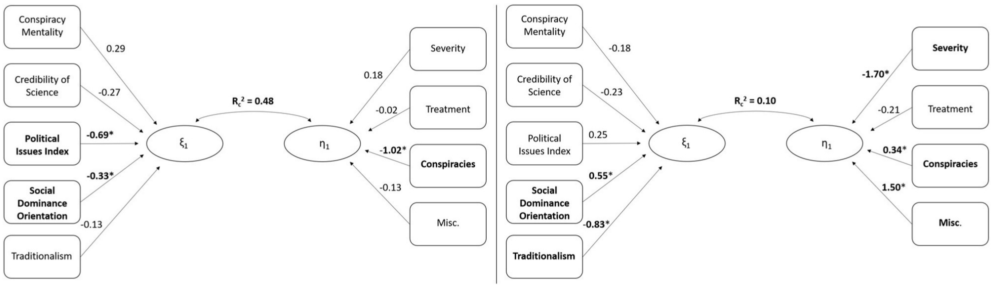


Figure 2.1: Diagram of the two significant canonical models. Left: More alignment with liberal policy positions and a low Social Dominance Orientation predict a low willingness to share conspiracy theories about COVID-19 on social media. Right: A high Social Dominance Orientation and a low endorsement of traditionalism predict a low willingness to share misinformation on social media related to the severity and spread of COVID-19, but a high willingness to share conspiracies about COVID-19 and miscellaneous cultural misinformation about COVID-19.

validate the patterns of individual differences and misinformation-sharing inclinations reported here.

2.3 Comparing Network Structures Across Different Kinds of Viral COVID (Mis)Information

2.3.1 Motivation

While there are a variety of ways in which one could classify types of misinformation, relevant for the study, one which way is between conspiratorial misinformation and non-conspiratorial misinformation. Conspiracy theories refer to wildly speculative explanations of large-scale events that contradict official explanations and posit, without unambiguous supporting evidence, the existence of a malevolent and powerful group of actors as the causal force behind said event or state of the world (Douglas et al., 2019). In the context of COVID-19, there are a wide variety of conspiracy theory allegations, including but not limited to: allegations that the pandemic itself is a hoax, claims that the virus is a genetically engineered bioweapon, and claims about how the pandemic is being used as a tool by malevolent individuals and/or groups to justify the taking away of civil liberties (van Mulukom et al., 2022). In contrast, much COVID-19 misinformation falls under the category of all other false information claims that do not explicitly or directly support any conspiracy theory. Much of this category involves false information about prospective treatments or preventative measures for COVID-19, and would likely be categorized as "junk science" misinformation in a taxonomy like that used by Carrasco-Farré (Carrasco-Farré, 2022), or as "cure, prevention, & treatment" misinformation by a taxonomy used by Charquero-Ballester and colleagues (Charquero-Ballester et al., 2021).

Just as different kinds of misinformation found online have been shown to have different structural features, so too are there differences in how misinformation diffuses across online networks and how people engage with misinformation. Prior to the COVID-19 pandemic, research by Vosoughi and colleagues (Vosoughi, Roy, & Aral, 2018) found that false information on the microblogging site formally known as Twitter diffused across larger networks than either true and mixed information did and did so at a faster rate. In another study examining the network structure of information and misinformation dissemination about the Zika virus on Twitter, misinformation networks were more likely than real information networks to evince more user-to-user dissemination and more involvement of users with disproportionately high influence as measured by out-degree centrality (Safarnejad et al., 2020).

Additionally, research analyzing trends in Google and Instagram search terms found that terms associated with COVID-19 conspiracies (e.g. about a link between COVID-19 and 5G technology or Bill Gates) were among the most searched for COVID-19 related terms on those platforms, and that searches for terms associated with cures for COVID-19 spiked following remarks by Donald Trump, then occupying the seat of the US President, about miracle cures and injecting disinfectants as ways to combat COVID-19 (Rovetta & Bhagavathula, 2020).

2.3.2 Approach

It has been common for researchers examining COVID-19 information and misinformation on social media to collect and analyze large data sets of posts for analyzing the transmission of misinformation across social media platforms. In contrast, our study focuses on a smaller number of posts, from the microblogging website formally known as Twitter, that received at least a moderately high amount of engagement. On this platform, and common to other microblogging sites such as BlueSky, Threads, and Mastodon, posts can be shared by other users and the frequency with which posts are shared are a metric to assess the virality of the post. From this smaller number of high engagement posts, we aimed to examine the associated follower networks of users sharing the original post. This is an effort to look more in-depth for network characteristics that could potentially be used to distinguish between networks engaging with information, conspiracy theory networks, and other kinds of non-conspiratorial misinformation networks. Being able to identify the nature of the misinformation being disseminated may aid in developing content- or context-specific strategies for combating the spread of misinformation.

2.3.3 Data

We restricted our study to individual posts that had gained a high degree of attention from users on the platform. For our purposes, we considered these posts to be those that had been shared a large number of times. Thus, our study explores the follower network structure of viral information posts, conspiratorial misinformation posts, and non-conspiratorial misinformation posts about COVID-19. Because these posts had each individually garnered high engagement, but the content differs substantively, it is possible that there are markers of different kinds of engagement or subcommunity structure in the follower/following networks of people who share viral (mis)information post. For example, we might hypothesize larger but more tightly clustered subcommunities engaging with COVID-19 conspiracies than with non-conspiratorial prevention or treatment misinformation, given prior research has found a negative relationship between accepting COVID-19 conspiracy theories and accepting alleged health protective information (Allington et al., 2021; Bridgman et al., 2020; van Mulukom, 2020).

For this study, we scraped tweets from Twitter, back when it was still known as Twitter. We selected tweets conveying either one of two different kinds of COVID-19 misinformation, conspiratorial and non-conspiratorial, or verified information. The content of the tweets selected for this study is shown in the Supplemental Material. To add an extra layer of user protection, tweet content is not provided verbatim, but in paraphrased form. Using Twitter's search function, we searched for tweets about the COVID-19 pandemic (using terms: `COVID-19`, `COVID`, `COVID19`, `coronavirus`, or `SARS-CoV-2`) that had been retweeted 1,000 and 15,000 times. One author [either MP or EJCL] initially evaluated each tweet to first determine if it was conveying misinformation. The subset of tweets that passed this initial screening were then discussed by all three authors to determine if they could be cleanly categorized as conveying conspiratorial misinformation, non-conspiratorial misinformation, or verified information. The conspiracy misinformation tweets promoted unverified or false theories about the origin of the SARS-CoV-2 virus (that it was deliberately created in a laboratory, government sanctioned, created as a bioweapon, etc.) as well as allegations that governments are misreporting COVID-19 deaths or data, and that COVID-19 cures exist that governments are hoarding. In contrast, the non-conspiratorial misinformation tweets contained advice about how to cure or prevent the transmission of COVID-19 that is unsupported by relevant medical research (homemade herbal recipes for cures, vitamin regimens, etc.). The tweets conveying verified information were predominantly about reported case rates or deaths associated with the then-novel coronavirus, but also included a tweet announcing a U.S. Justice Department decision allowing employers to mandate vaccines for employees to return to work.

As stated above, we limited our tweet selection to include only tweets with between 1,000 and 15,000 retweets. We selected the upper limit for computational feasibility, and to allow a network structure to emerge while maintaining interpretability. We selected the lower limit to ensure that there were a sufficient number of users to reveal meaningful network structure and that the misinformation or information was diffusing throughout a sufficient amount of users. The lower limit unfortunately has the side effect of drastically reducing the number of potential tweets for inclusion in this study, as only a small fraction of tweets garner at least that level of engagement. Previous works suggest that approximately 5% of tweets receive at least 1000 retweets (Nesi et al., 2018). The manual selection process for inclusion in the study is due to the specific focus in this study on comparing conspiratorial misinformation, non-conspiratorial misinformation, and verified information. Many high engagement tweets about COVID-19 were not conveying information *per se*. Instead, many high engagement tweets referencing COVID-19 were making jokes (e.g. a tweet expressing sadness at no longer being able to cancel plans due to COVID and having to restart using their usual bag of lies), were drawing attention to other non-COVID issues (e.g. a tweet expressing the dilemma facing U.S. high school teachers of opting to have classroom doors open for ventilation or classroom doors closed to increase safety from school shooters), or were issuing a warning (e.g. a tweet warning about a surge in cases following large numbers of students going to typical Spring Break destinations). Only by manually searching through the high engagement tweets fitting our inclusion criteria allowed us to determine whether the content of the tweet is conveying conspiratorial or non-conspiratorial misinformation.

It is worth remarking here that several high engagement tweets conveying misinformation were not included in the analysis because the contents of such tweets did not allow them to be cleanly categorized as conspiratorial or non-conspiratorial in nature, such as tweets regarding the prospects of ivermectin or hydroxychloroquine as a viable treatment for COVID-19. Likewise, we further excluded from consideration tweets that were about COVID vaccine safety or efficacy. In the past few decades, the topic of vaccines generally has been one in which there are active disinformation campaigns around, blending aspects of science denial, pseudoscience promotion, and conspiracy narratives (Mabrey, 2021). By limiting our study to tweets in this way, we hoped to limit the confounding nature of vaccine misinformation that, absent certain extra context (e.g. the source of the misinformation), would make it difficult to confidently code any misinformation as conspiratorial or non-conspiratorial.

Table 2.1: Table of retweet network statistics (note that edge numbers reflect the directed graphs).

Metric	Non-Conspiratorial	Conspiratorial	Verified Information
Minimum Node Count	*1515*	1787	2297
Maximum Node Count	3117	*9049*	14652
Average Node Count	2415.7 (475.57)	3892.8 (1945.0)	6237.63 (4049.62))
Minimum Edge Count	*2673*	6576	16492
Maximum Edge Count	36147	*282564*	258449
Average Edge Count	18445.2 (10964.5)	*62833.2* (75963.5)	97727.1 (77734.66)

To obtain retweets we applied, and were approved, for the Academic Twitter API. Using an interface called Postman, we scraped retweets (not including quote tweets) by including the body of the tweet with the `is:retweet` flag or the `retweets_of_tweet_id` flag. We then iterated through each page using the `next_token` for each individual search and saving retweets in the form of JSON files (in batches of 100 or 500). We then saved `user_id` information from each retweet. Using the list of obtained `user_id`'s, we scraped follower lists for each user using Tweepy and stored information in the form of a dictionary (where keys are users and values are lists of respective followers).[2] Once we obtained this information, we created a directed graph for each individual tweet. The creation of the graph is discussed in more detail below. A statistical summary of the selected tweets can be found in Table 2.1. From Table 2.1, we can see that, in general, the conspiratorial tweet and verified information retweet follower networks are larger. However, we note that there is large variance in the nodes and edges of the graph.

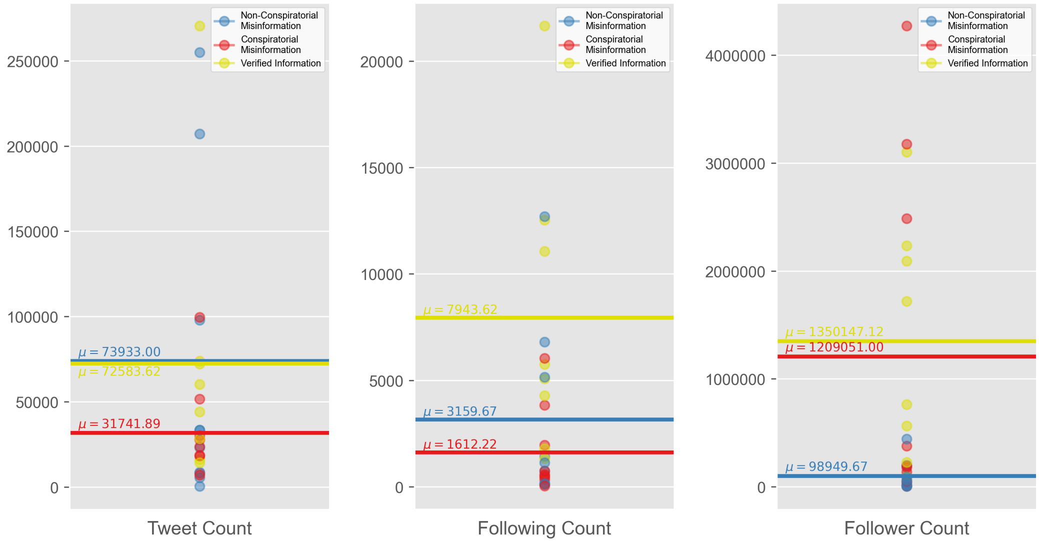


Figure 2.2: Summary statistics (tweet count, number of users following, number of followers) for tweet authors with averages for each category marked.

In addition to understanding the retweet network statistics, we also examine information about the original tweeters themselves. Figure 2.2 displays the tweet count (left), following count (middle) and follower count (right) for each of the eight to ten users that generated the semi-viral tweet. In general, nonconspiratorial and verified information users had higher numbers of tweets, while conspiratorial and verified information users had higher numbers of followers.

2.3.4 Methods

The desired information can be constructed as a directed graph. Each node in the graph represents a Twitter user. There is a directed edge between user A and user B if user A follows user B. Note that there can be a double-sided edge between user A and user B if they both follow each other. Social networking graphs, such as those generated by Twitter, can be extremely large and complex, making simple analyses difficult to perform. To understand the

[2] At this point, we inserted original tweet author information.

network structure of these viral misinformation tweets, we calculated centrality measures and carried out community detection analyses.

Centrality measures are an attempt to quantify the importance of each node in a network relative to the others. Some examples of centrality measures include in-degree (the number of directed edges into each node, referring to the number of users that follow this node) and out-degree (the number of directed edges out of each node, referring to the number of users the node is following). For the in-degree and out-degree metrics, we computed both the mean and the maximum. Other metrics we investigated included the percentages of unconnected nodes. More specifically, we measured unconnected in-nodes (percentage of users that were not followed by any of the other users) and unconnected out-nodes (percentage of users that did not follow any of the other users). In addition to these degree-based measures, we also examined betweenness, which measures how often each node is on the shortest path between two users. For a recent review on identification of influential nodes in Twitter, see (Riquelme & González-Cantergiani, 2016) and references therein.

Beyond using metrics to determine the characteristics of influential nodes, it is of interest to be able to detect the number of communities present in these networks. Community detection has been an active area of research involving statistical inference, machine learning, or other mathematical techniques. For reviews of methods for community detection, both classical and machine learning-based, see (Malliaros & Vazirgiannis, 2013; Fortunato & Hric, 2016; Traag, Waltman, & Van Eck, 2019; Jin et al., 2021) and the references therein. Similarly to influential node detection, community detection is also hindered by the understanding of what really defines a community (Fortunato & Hric, 2016). As community detection algorithms for directed graphs are still an active area of research, the communities measured in this manuscript are for the equivalent *undirected* graphs. In the context of this manuscript, we will use the Louvain (Blondel et al., 2008) algorithm for community detection. We note that the Louvain algorithm is intended for use in un-directed graphs. Thus, for our calculations of communities and modularity, we used a simplification of our graph in which edges were not directed. We assume a connection in the undirected graph if there is at least one edge between two nodes. We note that we also repeated the modularity calculations for the undirected network where an edge was implemented if there were two edges between two nodes (so called our bi-directional undirected network). We are interested in comparing the size and connectivity of the communities and observing whether or not there are differences between the communities in conspiratorial versus non-conspiratorial misinformation. In addition to detecting the communities, we also measured the modularity of each graphs using our detected communities (Blondel et al., 2008). Modularity is a measure between -1 and 1, where positive values signify that there is community structure present in the partition of communities present in the graph. Higher values of modularity indicate that the communities detected by our community detection algorithms are more densely connected with fewer connections between sub-communities. Modularity is not a feature inherent to the graph itself, but is calculated using community partitions. Since modularity can be sensitive to the community detection algorithm used, we implemented three different community detection algorithms and compared the measured modularity. The community detection algorithms employed included the Louvain method (Blondel et al., 2008), the Clauset-Newman-Moore greedy modularity maximization (Clauset, Newman, & Moore, 2004), and the Leiden Algorithm (Traag, Waltman, & Van Eck, 2019). In the Tables, we report the results generated by the Louvain method, but note that we found very little variability in modularity measurements between the 3 community detection algorithms (< 0.03).

2.3.5 Results

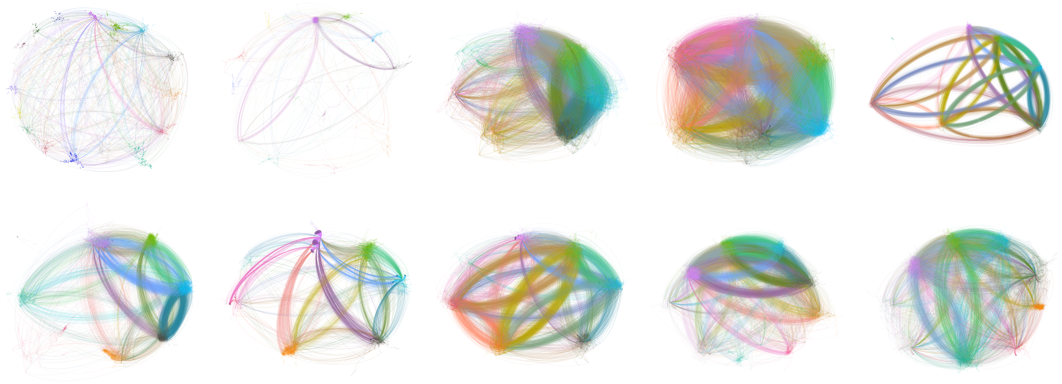


Figure 2.3: Network graphs for the retweeters for each non-conspiratorial tweet, where nodes and edges are colored by community membership (largest communities in purple, followed by light green, turquoise, and black). Sizes of the nodes are based on the in-degree of each node. Communities are only plotted if they comprise $> 1\%$ of the total retweet network.

Here we present the results from the analysis of the created tweet graphs. We note that, as this is an exploratory study, we do not have a sufficient sample size to conduct meaningful inferential statistical analyses or perform any machine learning techniques for classification. This exploratory study is meant to investigate whether there may be differences in network structures and to provide an idea of what types of metrics may be useful in distinguishing types of misinformation.

We begin with the graph structure itself. Figure 2.3, Figure 2.4, and Figure 2.5 display the 10 non-conspiratorial graphs, the 10 conspiratorial graphs, and the 8 information graphs, respectively. The colors within each graph signify the various detected community membership. The color of the edges between the nodes signifies the median color between the two nodes that are connected. For example, if a purple node is following a black node, the edge will be a darker purple. The graphs were constructed using the open-source library Gephi (Bastian, Heymann, & Jacomy, 2009). The communities were obtained using the default Gephi community detection algorithm based on the Louvain method (Blondel et al., 2008). The Louvain method was performed on the undirected version of our graph in which connections exist if there is at least one edge connecting two nodes. The largest communities are in purple, followed by light green, turquoise, and black. The largest communities (purple) are placed at the top of each graph and communities decrease in size as we rotate clockwise. In these graphs, we only plot the community if it comprises $> 1\%$ of the total retweet population. Thus, nodes that do not belong to a community that is 1% of the population or larger are not plotted in Figure 2.3, Figure 2.4, or Figure 2.5. While these figures generally show the large majority of the retweeters (85% total retweeters), there are some exceptions (Nonconspiratorial Tweets 1 and 2, as well as Conspiracy Tweet 2).

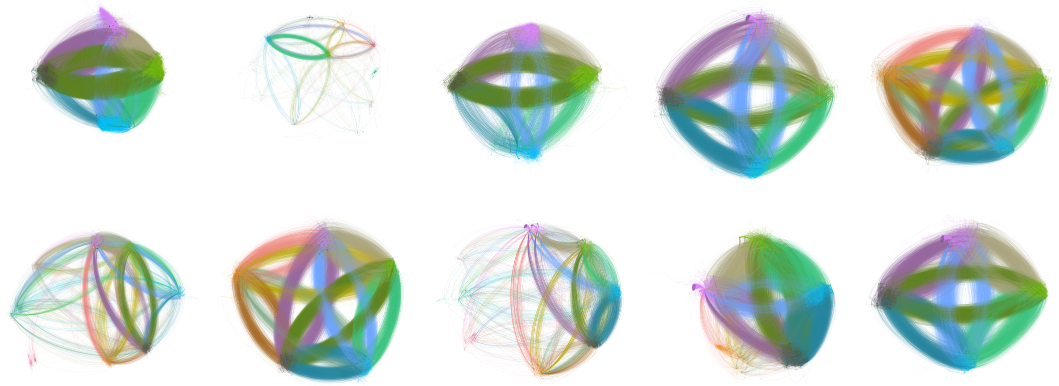


Figure 2.4: Network graphs for each conspiratorial tweet, where nodes and edges are colored by community membership (largest communities in purple, followed by light green, turquoise, and black). Sizes of the nodes are based on the in-degree of each node. Communities are only plotted if they comprise $> 1\%$ of the total retweet network.

When comparing the retweet graphs of the non-conspiratorial information in Figure 2.3 with the retweet graphs of conspiratorial information in Figure 2.4, it is clear that, in general, these conspiratorial graphs feature fewer, larger communities that are more densely connected. The majority of conspiratorial graphs feature 5 or fewer communities, while the smallest number of communities detected for a non-conspiratorial network is 6. When comparing the follower graphs of re-tweeters in the verified information tweets, we also see, in general, fewer and more interconnected communities compared to the non-conspiratorial misinformation tweets.

In addition to looking at the graphs of the retweeters, we can also examine whether or not there are differences in the timing of the retweets between conspiratorial, non-conspiratorial, and verified information tweets. Figure 2.6 displays the percentage of retweets over the first three days after each original tweet was created. The top panel contains all of the tweets colored by non-

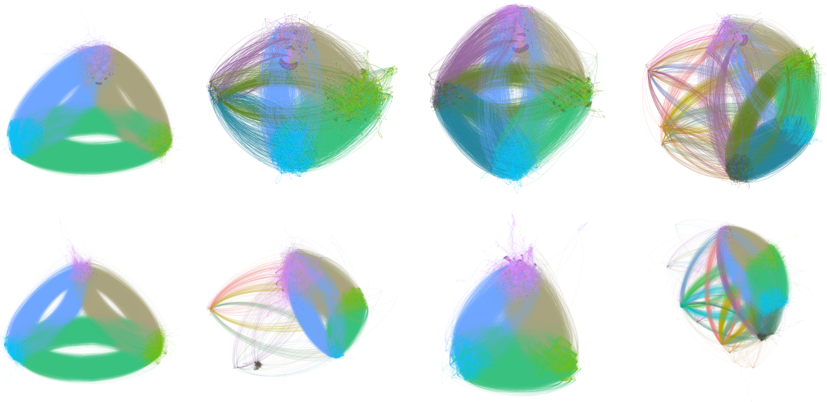


Figure 2.5: Network graphs for each information tweet, where nodes and edges are colored by community membership (largest communities in purple, followed by light green, turquoise, and black). Sizes of the nodes are based on the in-degree of each node. Communities are only plotted if they comprise $> 1\%$ of the total retweet network.

conspiratorial (blue), conspiracy (red), and information (yellow). The bottom panel examines the average over the non-conspiratorial, conspiracy, and information tweets categories. We observe there are no large differences between the timing of retweets between the three categories of information, but find that the initial spike and eventual decay is consistent with a typical retweet trend (Kobayashi & Lambiotte, 2016). Although we cannot make any statistical differences, we do notice that the misinformation tweets, on average, have a secondary bump between 6-12 hours which is not observed in verified information tweets.

2.3.6 Graph Analysis

We calculated the major centrality measures and community detection for the full retweet graphs for the non-conspiratorial, conspiratorial misinformation, and verified information tweets (i.e. including the original tweeter). Table 2.2 portrays the information for the various metrics. Each metric is reported with means (standard deviations) for the tweets used in the analysis.

The in-degree measurements vary greatly between the non-conspiratorial misinformation and both verified information and conspiratorial misinformation. This indicates that the conspiratorial and verified information networks may be more inter-connected than the non-conspiratorial networks. This is supported by the fact that non-conspiratorial misinformation graph partitions have higher modularity than the other two types. Graph partitions with high modularity signify subcommunities that have dense connections within the subcommunity but sparse connections between subcommunities. When examining the size of the largest communities, we observe that the largest communities in the conspiratorial misinformation and verified

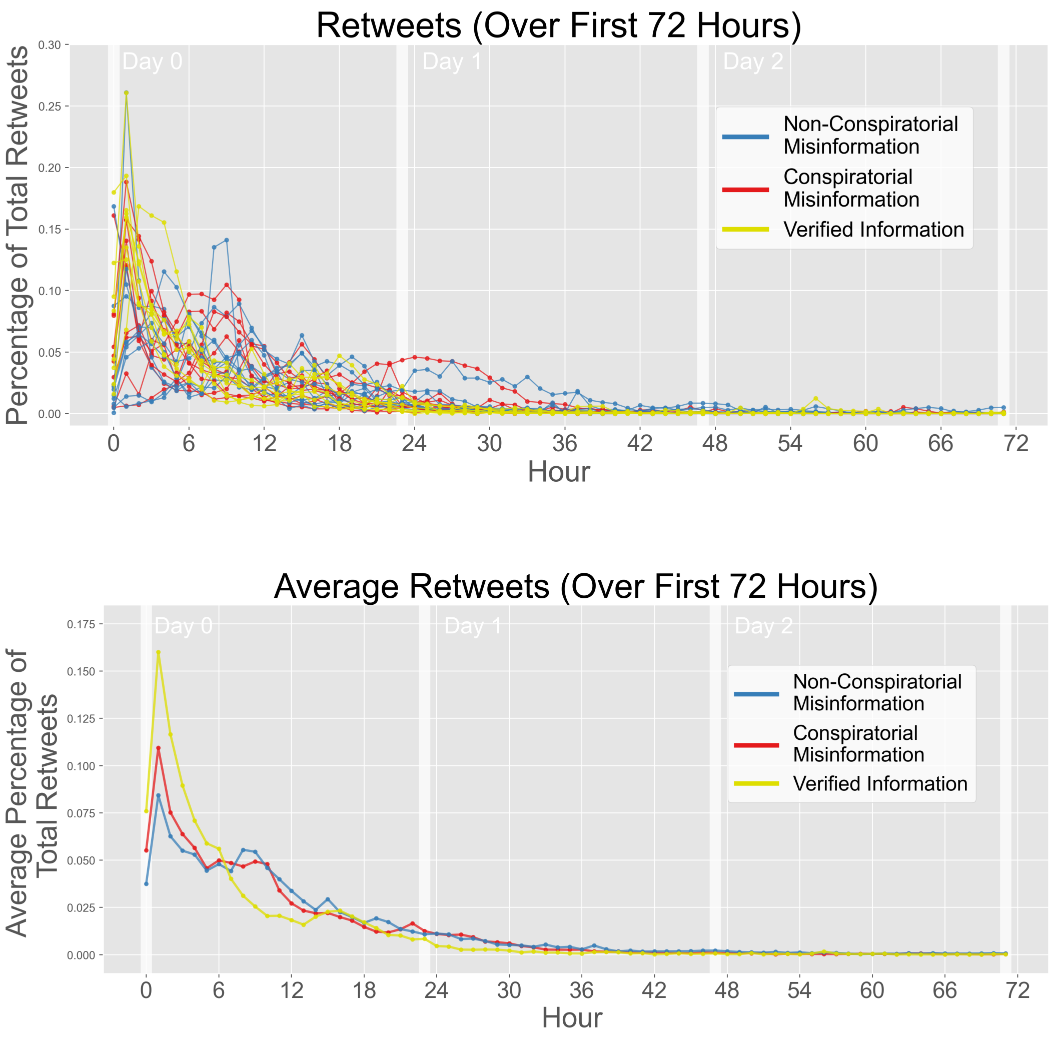


Figure 2.6: Non-conspiratorial, conspiratorial, and verified information retweets over time. In the top figure, percentage of total retweets for each tweet are plotted over the first 72 hours after tweet creation. In the bottom figure, percentage of total retweets per category of tweet are averaged over the first 72 hours after tweet creation.

information networks were substantially larger than the non-conspiratorial misinformation networks (36.8% for conspiratorial, 48.5% for verified information versus 24.5% for non-conspiratorial misinformation). Figure 2.7 shows the percentage of the network in each largest community for all tweets. Overall, we can see that conspiratorial misinformation and verified information tweets tend to have larger communities than the non-conspiratorial misinformation tweets. Moreover, when counting the number of communities that comprised at least 1% of the retweet network, we found that there were fewer communities for conspiratorial misinformation (5.9) and verified

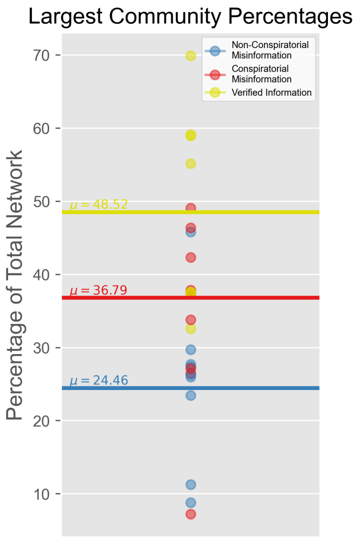


Figure 2.7: Percentage of nodes belonging to the largest community for non-conspiratorial misinformation (blue) vs. conspiratorial misinformation (red) vs. verified information (yellow) networks with averages marked.

information (4.6) than nonconspiratorial misinformation (9.4) tweets.

2.3.7 Discussion

In this study, we compared the network structure of viral microblogging posts that contained either verified information, conspiratorial misinformation, or non-conspiratorial misinformation about the COVID-19 pandemic to examine how categories of (mis)information are engaged with by social media users. Broadly, our results revealed that high-engagement posts expressing conspiratorial COVID-19 misinformation had fewer but larger sub-communities engaging with the content compared to high-engagement posts expressing non-conspiratorial COVID-19 misinformation. This was revealed through community detection analysis, where, on average, there were 5.9 detected communities that contained more than 1% of a given conspiracy tweet's network, as compared to 9.4 detected communities of at least 1% of the non-conspiratorial misinformation tweets' network. This can also be observed in the average size of the largest communities for conspiracy tweets accounting for a greater percentage of the network (μ = 36.79%) compared to the percentage of the network captured in the largest communities detected for the non-conspiratorial misinformation tweets (μ = 24.46%). Of note, the network structure for verified information more closely resembled conspiratorial misinformation networks than non-conspiratorial misinformation networks, with 4.6 detected communities containing more than 1% of a given post's network. Additionally, the average size of the largest communities detected in the verified information networks accounted for 48.52% of the total network, again

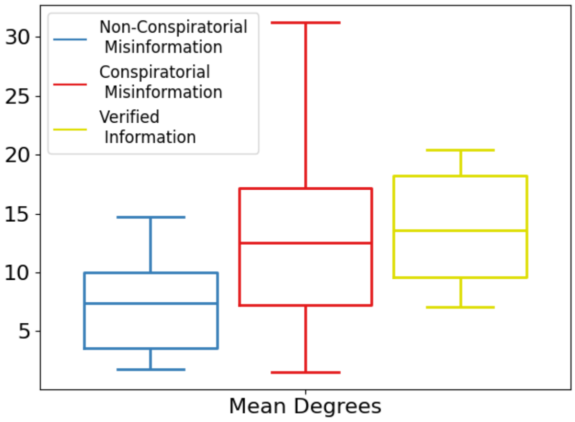


Figure 2.8: Boxplot of the degrees for both the non-conspiratorial misinformation (blue), conspiratorial misinformation (red), and verified information (yellow) tweets.

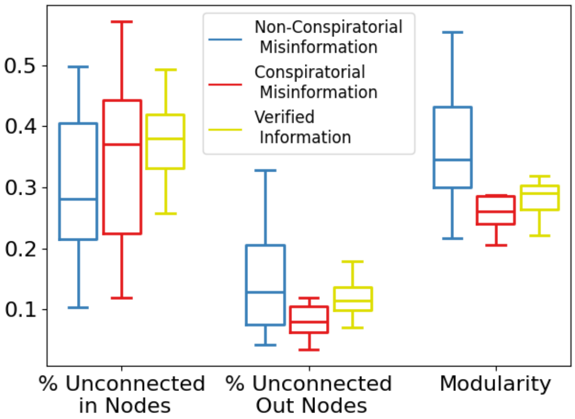


Figure 2.9: Boxplot of various metrics for non-conspiratorial misinformation (blue), conspiratorial misinformation (red), and verified information (yellow) tweets, including percentage of unconnected in-nodes (left), unconnected out-nodes (middle), and modularity (right).

Table 2.2: Table of mean centrality values for follower-followee graphs retweeters of non-conspiratorial misinformation, conspiratorial misinformation, and verified information. Values are presented as means (standard deviations) for each category.

Metric	Non-conspiratorial Misinformation	Conspiratorial Misinformation	Verified Information
In-Degree	7.61 (4.42)	13.599 (8.189)	13.723 (4.974)
% Unconnected Nodes (in)	32.79 (18.28)	33.90 (13.74)	37.56 (6.78)
% Unconnected Nodes (out)	15.13 (9.78)	11.94 (12.73)	11.824 (3.27)
Closeness	0.1463 (0.08646)	0.1641 (0.07213)	0.1375 (0.02549)
Betweenness (10^{-4})	6.091 (2.220)	4.382 (2.439)	3.365 (2.057)
PageRank (10^{-4})	4.328 (0.9827)	3.087 (1.213)	2.259 (1.248)
Modularity	0.3942 (0.1439)	0.3011 (0.1155)	0.283 (0.02974)
Largest community (%)	24.46 (10.37)	36.79 (14.53)	48.52 (13.84)
# of communities ($> 1\%$)	9.400 (4.195)	5.900 (2.846)	4.625 (2.066)

more closely resembling conspiratorial network structure than non-conspiratorial misinformation network structure. Although we lacked a sufficient sample size of tweets to produce meaningfully interpretable output from inferential statistical analyses, descriptively we observed less variation in the percentage of unconnected in nodes and out nodes for the sets of conspiracy and verified information networks than for the set of non-conspiracy misinformation networks. Likewise, the non-conspiratorial misinformation networks had overall lower and less varied degree centrality scores than the verified information and conspiracy networks. This suggests that there are more highly interconnected communities engaging with either verified information or conspiratorial posts than for non-conspiratorial misinformation about COVID-19.

In the context of COVID-19, at least, these features might indicate the presence of echo chambers, or closed epistemic communities reinforcing and validating a community's attitudes or beliefs through repetition (Cinelli et al., 2021). This may manifest as an echo chamber validating the supposed truth behind the conspiratorial claim, or even as an echo chamber validating the mockery and satire directed towards the absurdity of the conspiratorial claim. Regarding communities engaging with verified information, the echo chambers may be in service of validating acceptance or rejection of the verified information. In either case, the relative lack of "cross-talk" between communities in the network structure for both verified information networks and conspiratorial misinformation networks contrasts with what we observed regarding non-conspiratorial misinformation networks. This kind of misinformation spread to a larger number of communities than the conspiratorial misinformation did. Regarding non-conspiratorial misinformation, this finding aligns with the findings of Röchert and colleagues (Röchert et al., 2021) that users who disseminated COVID-19 misinformation on YouTube were connected to heterogeneous networks, which faciliates the spread of misinformation across many distinct communities. However, it should be noted that these researchers did not distinguish between conspiratorial misinformation and non-conspiratorial misinformation in their study.

The difference in network structure between the conspiratorial and non-conspiratorial viral tweets may relate to the intended purpose behind these different kinds of false claims. The non-conspiratorial misinformation conveyed in the tweets we scraped may be more likely to have emerged as an empirical claim whose alleged truth value was intended to help individuals navigate the pandemic during times of heightened uncertainty. While this does not discount the possibility that such false information may have been intentionally fabricated or disseminated for manipulative purposes – to enhance the esteem of its originator or to sell products, for instance – the nature of the claims themselves can be acted on in direct ways. This may facilitate their spread across communities. By comparison, the conspiratorial false information conveyed in the conspiracy tweets we scraped does not directly lend itself to individual action. Likewise, the verified information tweets tended to convey updates on the pandemic rather than to communicate any actionable messaging. Thus, rather than serving to help individuals successfully navigate the world, engaging with conspiratorial misinformation or verified information may serve as a social signal, with how one engages with the information (e.g. endorsing or refuting) conveying an individual's broader ideological allegiances. This interpretation aligns both with theorizing as to multiple distinct functions for different kinds of beliefs generally (Funkhouser, 2017) as well as theorizing about conspiracy beliefs more specifically (van Prooijen & Douglas, 2018). Beliefs, even false beliefs, intended to facilitate successful navigation may simply be more likely to spread across more communities of people than beliefs (independent of their truth value) that serve to signal, validate, or reinforce social status information.

Our results contribute to efforts aiming to understand the nature of different kinds of misinformation by identifying network feature differences in conspiratorial and non-conspiratorial COVID-19 misinformation shared over microblogging sites. All of the misinformation conveyed in the tweets we scraped have the potential to be dangerous, albeit in different ways. Non-conspiratorial health misinformation may encourage people to engage in behaviors that are not medically recommended or may discourage people from engaging in medically recommended behaviors (McGlynn, Baryshevtsev, & Dayton, 2020). As such, while this misinformation is still potentially dangerous, the potential is different than the danger posed by the spread and endorsement of conspiratorial misinformation. Conspiratorial misinformation carries with it the additional potential for dangerous behaviors on a longer-term and greater-scale. Prior research on conspiracy belief acceptance has shown it to be related to a reduced willingness to engage in health-protective behaviors as well as to an increased endorsement of political violence and overall political disengagement, both within and outside the context of COVID-19 (van Mulukom et al., 2022; Vegetti & Littvay, 2021; Lamberty & Leiser, 2019). Further, our results suggest that efforts to recognize conspiratorial misinformation on the basis of network structure may be hampered by the similarity with which networks of users appear to engage with both conspiratorial misinformation and verified information.

As such, even though there are numerous research efforts showing substantial covariation between endorsing conspiratorial and non-conspiratorial misinformation (Lewandowsky, Oberauer, & Gignac, 2013; E. Lobato et al., 2014; Rizeq, Flora, & Toplak, 2021), recognizing distinct patterns in how different kinds of misinformation spread is useful for researchers who study misinformation. Not all misinformation is created equal (Carrasco-Farré, 2022), nor do people seem to engage with different kinds of misinformation equally. Despite some structural similarities between conspiratorial and non-conspiratorial misinformation about the COVID-19 pandemic, our results reveal potential differences in the network structure between these two broad kinds of misinformation. It is therefore beneficial for misinformation researchers to attend to the specific nature of the misinformation they are investigating beyond its truth value, as it may be the case that successfully combating the spread of different kinds of misinformation will require different approaches.

Chapter 3

Socio-political Movements

3.1 Introduction

Hashtags are a fundamental feature of Twitter, playing a pivotal role in organizing content, facilitating discoverability, and fostering engagement within the platform's vast ecosystem. Essentially, hashtags are keywords or phrases preceded by the pound (#) symbol, which categorize tweets and enable users to participate in broader conversations centered around specific topics, events, or themes. When users search for or click on a hashtag, Twitter displays a feed of tweets containing that particular tag, allowing individuals to explore a range of related content. This mechanism not only aids in content discovery but also fosters community engagement as users contribute to ongoing discussions or trends associated with the hashtag. Moreover, hashtags serve as powerful tools for organizing information and generating real-time insights (Adamska, 2015).

During significant events such as protests or breaking news stories, hashtags become virtual meeting points where users converge to share updates, opinions, and reactions. This collective aggregation of tweets under a common hashtag enables users to stay informed, express solidarity, or voice dissent, effectively transforming Twitter into a dynamic hub for social discourse and activism. Select hashtags on Twitter emerge as calls to action or to bring awareness/attention to acts of injustice, often leading to the mobilization of individuals from all walks of life in an effort to incite change. These hashtags define the political climate at the time and become hot topics that are debated on any platform imaginable (on the news, on social media, amongst friend groups, etc.). These viral social hashtags underscore the power of social media as a catalyst for social change, providing a platform for marginalized voices, fostering community solidarity, and driving conversations about pressing social issues. They demonstrate how digital activism can translate into real-world impact, prompting individuals and institutions to reckon with systemic injustices and work towards a more equitable and inclusive society (Goswami, 2018).

Among the most widespread and influential socio-political hashtags that have emerged in recent years is #BlackLivesMatter, originating in the United States and eventually gaining global prominence. Its inception came with the murders of Trayvon Martin and Michael Brown and subsequent lack of criminal convictions for their killers in 2013 and 2015, respectively (Garza, 2014; Francis & Wright-Rigueur, 2021). The hashtag later evolved to bring awareness to many other acts of injustice against Black members of the population, primarily by police, combating systemic racism, and advocating for the rights and dignity of Black people. Throughout its evolution, the Black Lives Matter movement has emphasized the importance of centering the experiences and voices of Black communities, challenging institutions and policies that perpetuate racial inequality, and advocating for tangible reforms to create a more just and equitable society.

In response, the hashtag #AllLivesMatter was created to assert “colorblind" attitudes ostensibly at odds with sentiments expressed by #BlackLivesMatter (Orbe, 2015; Ince, Rojas, & Davis, 2017; Tawa, Ma, & Katsumoto, 2016; Gallagher et al., 2018). The #AllLivesMatter hashtag emerged as a response to the Black Lives Matter movement, aiming to assert that all lives have value regardless of race or ethnicity. While the phrase “All Lives Matter" may seem to assert that all lives have value regardless of race or ethnicity, its usage often undermines the specific focus on addressing systemic racism and inequality faced by Black communities. Critics argue that deploying this hashtag dilutes the urgency of addressing the unique challenges and injustices experienced by Black people, deflecting attention away from the need for systemic reforms. Moreover, the #AllLivesMatter hashtag has been criticized for perpetuating a colorblind narrative that fails to acknowledge the historical and ongoing disparities faced by marginalized communities. Its usage in response to calls for racial justice has sparked debates about the nuances of solidarity, empathy, and the importance of centering marginalized voices in conversations about equity and social change.

3.2 Pairwise Hashtag Comparison

Collaborators: Alex John Quijano, Ayme Tomson, Prof. Arnold Kim, & Prof. Suzanne Sindi

3.2.1 Motivation

While many viral hashtags have emerged as powerful tools for social movements, they are often met with counter-hashtags that express diverging opinions. For instance, when a hashtag brings attention to a specific form of injustice or inequality, a response hashtag might surface to challenge the original message. These counter-hashtags aim to shift the conversation, sometimes focusing on perceived negative consequences of the initial movement or promoting a broader, more “inclusive" viewpoint. This dynamic reflects the contentious and polarized nature of social discourse, revealing deep societal divides and the complexity of addressing widespread social issues.

A past study compared the divergent discourse around #BlackLivesMatter and #AllLivesMatter, and found that the Counter Hashtag (#AllLivesMatter) only introduced opposition and tension to the overall conversation, and that the Original Hashtag (#BlackLivesMatter) yielded more diverse conversations (Gallagher et al., 2018). We extend this study by using some of the same methodologies, as well as some new methods, on three more pairs of divergent hashtags in an effort to generalize the nature of online disagreement and classify hashtags based on characteristic conversation.

3.2.2 Data

In addition to #BlackLivesMatter, several viral social hashtags on Twitter have emerged and catalyzed social movements and were also met with a response hashtag. We analyzed four pairs of hashtags: (1) #BlackLivesMatter vs. #AllLivesMatter, (2) #MeToo vs. #HimToo, (3) #TakeAKnee [1] vs. #StandForTheFlag, and (5) #GunControlNow vs. #2ndAmendment. In each case there is the “Original" Hashtag, which emerges as a result to some widespread form of injustice. The Original Hashtag is then followed by the emergence of a “Counter" Hashtag, which arises as a response to the Original Hashtag in an effort to express a divergent opinion and often incite debate.

- The **#MeToo** movement (Hillstrom, 2018) gained widespread traction on social media platforms in October 2017. Originally created by activist Tarana Burke over a decade earlier, the movement resurfaced on Twitter when actress Alyssa Milano encouraged women to share their experiences of sexual harassment and assault using the hashtag #MeToo. This quickly evolved into a global phenomenon, with millions of individuals, predominantly women, sharing their stories of harassment, abuse, and misconduct. The movement exposed the prevalence of sexual violence and challenged societal attitudes and power structures that perpetuate such behavior. It prompted discussions about consent, accountability, and the need for cultural and institutional change to create safer environments for all individuals.

  – The **#HimToo** (Boyle & Rathnayake, 2020) hashtag emerged as a counter-narrative to the #MeToo movement, gaining traction as individuals used it to highlight concerns about false accusations of sexual misconduct against men. This shift in focus diluted the impact of the #MeToo movement and detracted from its primary goal of advocating for victims and driving systemic change. As #HimToo gained momentum, it often overshadowed stories of survivors and their experiences, placing greater emphasis on the perceived dangers of false allegations. Such narratives not only undermined the

[1] We took the hashtags #TakeAKnee and #TakeTheKnee together and treated as the same.

visibility of #MeToo but also perpetuated a culture of disbelief and victim-blaming, hindering progress toward addressing systemic issues of sexual misconduct and gender inequality.

- The **#TakeAKnee** (Duckett & Sacra, 2019) hashtag emerged in September 2016 in response to NFL quarterback Colin Kaepernick's decision to kneel during the national anthem before football games as a protest against racial injustice and police brutality. Kaepernick's peaceful protest ignited a nationwide debate about the intersection of race, patriotism, and free speech. Following Kaepernick's lead, athletes across various sports, including NBA star LeBron James and soccer player Megan Rapinoe, used their platforms to raise awareness about social justice issues. While the hashtag garnered both praise and criticism, it played a significant role in amplifying conversations about racial injustice, police accountability, and the role of activism in sports. It also spurred broader discussions about the ways in which individuals can use their platforms to effect social change and advocate for marginalized communities.

  - The **#StandForTheFlag** hashtag emerged prominently in response to the #TakeAKnee movement. The #StandForTheFlag hashtag expressed a divergent sentiment that viewed the act of kneeling as negative and disrespectful. This divergent sentiment manifested in strong negative reactions, with some calling for boycotts of NFL games, demanding that players be fined or suspended. Media personality Laura Ingraham, made a dismissive and silencing comment, “shut up and dribble" (Niven, 2021), to suggest that athletes, particularly those in the NBA, should refrain from expressing their political and social views. The hashtag #StandForTheFlag encapsulated a broader backlash against the perceived politicization of sports and the use of national symbols as platforms for protest. While #TakeAKnee aimed to highlight and address racial inequalities, #StandForTheFlag was criticized for dismissing the underlying issues that prompted the protests.

- **#GunControlNow** emerged as a rallying cry in the wake of numerous mass shootings in the United States, particularly gaining momentum after the tragic Parkland school shooting in 2018. Survivors, activists, and concerned citizens utilized the hashtag to call for immediate and comprehensive action to address the profound effects of gun violence. Rooted in a deep sense of urgency and frustration, the hashtag galvanized to advocate for stricter gun control laws and policies aimed at preventing future tragedies. #GunControlNow facilitated the organization and coordination of large-scale protests, such as the March for Our Lives, which drew hundreds of thousands of participants from across the country (Phillips, 2019). These demonstrations served as poignant displays of unity and resolve, amplifying the voices of survivors and victims' families while demanding accountability from lawmakers and elected officials. By highlighting the human cost of inaction and the devastating impact of gun violence on communities, the hashtag generated public pressure on policymakers to prioritize the enactment of effective gun control measures.

- The **#2ndAmendment** hashtag emerged as a response to the #GunControlNow movement, representing a counter-narrative that emphasized the constitutional right to bear arms. The proliferation of the #2ndAmendment hashtag perpetuated a polarized and contentious debate around gun control. Furthermore, the #2ndAmendment hashtag contributed to a culture of resistance against any proposed gun control measures, even those aimed at preventing mass shootings and protecting public safety. This resistance often came at the expense of prioritizing the lives and well-being of individuals affected by gun violence. By framing gun control as a threat to Second Amendment rights, the hashtag perpetuated a narrative that prioritized individual liberties over collective safety, exacerbating the challenges of enacting effective gun reform.

We scraped tweets using the Twitter search feature via a Python Scrapy module available on Github ("jonbakerfish/TweetScraper: TweetScraper is a simple crawler/spider for Twitter Search without using API", n.d.). In particular, we scraped for tweets from January 2014 to October 2019 – a timeline in which Twitter use skyrocketed and numerous social justice hashtags emerged. We then cleaned the tweets by lowercasing and removing usermentions, punctuation, hyperlinks, and stop words (stop words were removed for Entropy & Divergence analysis only). Such cleaning was done using the Python's regular expressions module and the NLTK module (Bird, Klein, & Loper, 2009). For consistency, we randomly sampled 0.001% of the original data 50 times, creating 50 sets of simple random samples for each hashtag.

3.2.3 Methods

In order to evaluate structural differences between conversation surrounding pairwise Twitter hashtags, we performed analyses on tweets in two ways: (1) statistically using measures of entropy and divergence, and (2) using word embeddings and principal component analysis.

Entropy & Divergence

The Effective Diversity of a text,

$$D = 2^H, \tag{3.1}$$

is measured using the Shannon entropy,

$$H = -\Sigma_{i=1}^{n} p_i \log_2 p_i, \tag{3.2}$$

which can be thought of as a quantity that measures diversity or “unpredictability" of a given body of text (corpus) with n unique words. Here p_i denotes the probability of a word i occurring in that text. A high Shannon entropy measure thus implies a diverse and unpredictable corpus while a smaller Shannon entropy measure indicates a uniform and predictable corpus. Using the Effective Diversity allows us to compare the diversities of two corpora in an effective and meaningful way.

The Jensen-Shannon Divergence (JSD) between two corpora P and Q is

$$D_{JS}(P||Q) = \pi_1 \Sigma_{i=1}^{n} D_{KL}(P||M) + \pi_2 \Sigma_{i=1}^{n} D_{KL}(Q||M) \tag{3.3}$$

where

$$D_{KL}(P||M) = \Sigma_{i=1}^{n} p_i \log_2 \frac{p_i}{m_i} \tag{3.4}$$

is the Kullback-Leibler divergence, allowing us to assess the distributional difference between P and the mixed distribution,

$$M = \pi_1 P + \pi_2 Q. \tag{3.5}$$

Here, π_1 and π_2 are constants that scale the corpora based upon size. The JSD is bounded between 0 and 1, where a measure of 0 indicates that the two corpora are the exact same, and a measure of 1 indicates that the two corpora have no words in common at all.

We can additionally measure an individual word’s contribution to the overall diversity using

$$D_{JS,i}(P||Q) = -m_i \log_2 m_i + \pi_1 p_i \log_2 p_i + \pi_2 q_i \log_2 q_i. \tag{3.6}$$

Bidirectional Encoder Representations from Transformers (BERT)

BERT is a publicly available pre-trained neural network model that uses a self-attention mechanism to map words into a vector space based on context. This attention mechanism relates different positions in a singular sequence to produce a single representation. As a transfer learning model, BERT has been used for many types of information extraction tasks on Twitter data, including its use to evaluate hate speech and offensive content on Twitter (Benballa, Collet, & Picot-Clemente, 2019; Mozafari, Farahbakhsh, & Crespi, 2019). The model is pre-trained on millions of books and online articles, but for this work, not fine-tuned. Since its release, BERT has inspired numerous variations and derivatives, such as RoBERTa (Robustly Optimized BERT Approach) and DistilBERT (a smaller, faster, and more efficient version that retains much of BERT’s power but is optimized for speed and resource usage). These models extend BERT’s capabilities, making sophisticated NLP accessible for a broader range of applications. We utilize a pre-trained model from cardiffnlp, pre-trained on political tweets (Face, 2024).

3.2.4 Results

From preliminary analysis, we find that much of the conversation between pairwise hashtags is the same. In fact, in each case, the conversation around the Counter Hashtag was found to be a subset of the corresponding Original Hashtag. As a result, we can infer that while each hashtag expresses a different sentiment, similar topics, diction, and syntax are present even in opposing hashtags.

We find that overall, all hashtag pairs do not diverge very much, as all measures are below 0.4 and though #MeToo vs. #HimToo fosters the most diverse conversations, they diverge very little (least off all hashtag pairs in this study).

Individual Hashtags	Mean Effective Diversity		Mean Jensen-Shannon Divergence
	Individual Hashtags	Mixed Distribution	
#BlackLivesMatter	2256.526245	2357.711178	0.220186
#AllLivesMatter	924.236989		
#MeToo	4620.905919	4643.494839	0.049868
#HimToo	358.094238		
#TakeAKnee	1089.366669	1134.548872	0.181235
#StandForTheFlag	202.637369		
#GunControlNow	1774.873041	2137.107321	0.347292
#2ndAmendment	1642.710857		

Words that contribute most to the overall diversity are often other related, viral hashtags. For example, #bluelivesmatter and #policelivesmatter were in the top three words contributing to diversity in the #BlackLivesMatter vs. #AllLivesMatter pair, and #standfortheanthem and #gunreformnow contribute largely to the diversities of #TakeAKnee vs. #StandForTheFlag and #GunControlNow vs. #2ndAmendment, respectively. Finally, hashtags such as #maga (“Make America Great Again", Donald Trump’s campaign slogan), #black, and #people contribute to most to multiple diversities. Thus these words transcend topic, and are often discussed and/or subjects of debate.

Hashtag Pairs	**Tokens ordered by greatest contribution score**
#BlackLivesMatter vs. #AllLivesMatter	bluelivesmatter, black, policelivesmatter, prolife, life, liberty, people, evil, worse, bless
#MeToo vs. #HimToo	dates, climate, radical, respects, current, false, accusations, feminists, vote, confirmkavanaughnow
#TakeAKnee vs. #StandForTheFlag	standfortheanthem, boycottnfl, usa, disrespect, maga, kneel, pray, respecttheflag, rich, nflboycott
#GunControlNow vs. #2ndAmendment	gun, gunreformnow, maga, prayers, rights, teaparty, gunviolence, marchforourlives, neveragain, thoughts

Table 3.1: This table shows the top ten select words corresponding to each Twitter hashtag pair with greatest contributions to the overall diversity of the mixed distribution. Results shown are for only one sample set.

A purely textual approach does not show significant distinction between hashtag pairs.

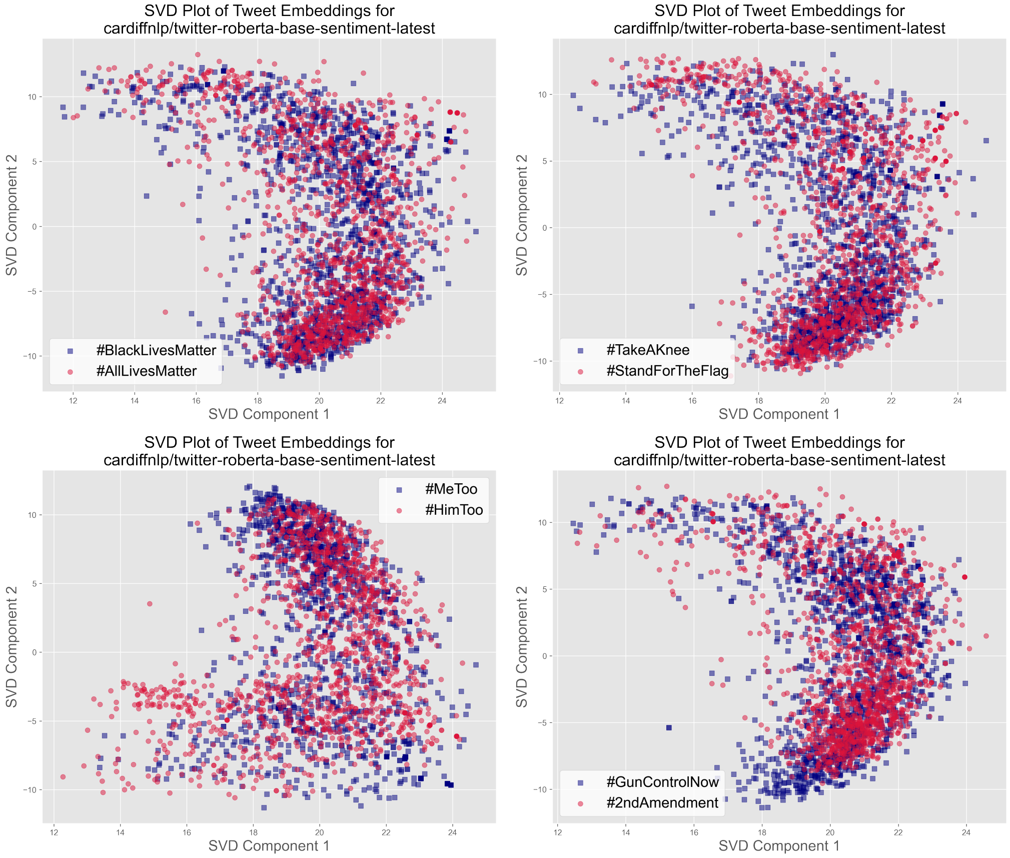


Figure 3.1: Results from performing Singular Value Decomposition on pre-trained BERT.

We thus conclude the measures of Entropy & Divergence and BERT word embeddings successfully characterize conversations around hashtags, but more work and/or tools are necessary for classification. Instead of fostering constructive dialogue, discourse between hashtag pairs often devolves into entrenched positions and ideological battles, with little room for compromise or nuanced discussion. However, understanding the nature of online disagreement can help to provide insight into how individuals communicate online, as well as how individuals navigate around sensitive topics in a volatile political climate.

3.3 Mutual Influence

3.3.1 Motivation

The relationship between the All Lives Matter (ALM) and Black Lives Matter (BLM) movements is complex and often contentious. One common criticism of the All Lives Matter response to Black Lives Matter is that it can be seen as dismissive or minimizing of the unique challenges and struggles faced by Black individuals and communities, failing to acknowledge

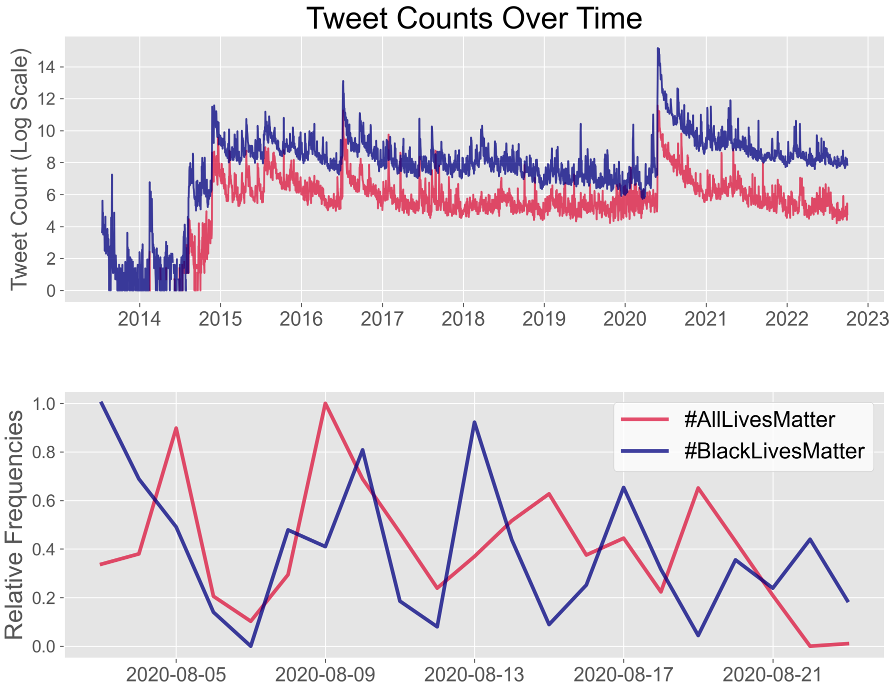


Figure 3.2: Sample interactions.

the specific historical and systemic injustices faced by Black communities. Supporters of Black Lives Matter argue that the movement is not about asserting that *only* Black lives matter, but rather, it is a call to address the systemic racism and violence disproportionately affecting Black people. Meaningful dialogue and reconciliation between these movements often require a deep understanding of the historical and social contexts that have shaped their respective perspectives on racial justice and equality.

The volume of tweets containing each respective hashtag naturally fluctuates over time, often peaking when the death of an individual is highly publicized. As such, there is lower engagement with each hashtag at times between viral tragedies. Potential exists for #BlackLivesMatter and #AllLivesMatter tweets to be related in more nuanced ways, both characteristically and, more notably, mathematically (i.e. does the increase in the use of one hashtag contribute to the increase or decrease of the other? Or vice versa?). To explore the potential interplay, we focus on three widely-covered events:

- **Sandra Bland**, a 28-year-old Black woman from Illinois, was pulled over for a routine traffic stop in Waller County, Texas, by a Texas State Trooper. The reason for the stop was reportedly a failure to signal a lane change. Video footage from the police dashboard camera captured the quickly escalating encounter, where Bland was forcibly removed from her vehicle. She was ultimately arrested and charged with assaulting a public servant, then detained at the Waller County Jail. Three days later, on July 13, 2015, Sandra Bland was allegedly found dead in her jail cell. Authorities claimed she died by suicide, hanging herself with a plastic trash bag. However, her family and supporters raised doubts about this narrative, suggesting foul play or negligence on the part of the authorities. Bland's death sparked widespread outrage and protests, with many questioning the circumstances surrounding her arrest and subsequent death. Bland's family filed a wrongful death lawsuit against the Texas Department of Public Safety and the Waller County Sheriff's Office, among others. The case resulted in a $1.9 million settlement for the family (Klein, 2018).

- **Breonna Taylor**, a 26-year-old Black woman, was fatally shot by police officers while she was sleeping during a botched raid on her apartment on March 13, 2020, in Louisville, Kentucky. The officers were executing a search warrant in connection with a narcotics investigation, but Taylor was not the target of the investigation, and no drugs were found in her home. In September 2020, the city of Louisville reached a $12 million settlement with Breonna Taylor's family in a wrongful death lawsuit. On September 23, 2020, a grand jury indicted one of the officers involved on three counts of wanton endangerment for firing shots that entered a neighboring apartment. However, none of the officers were directly charged in connection with Taylor's death, leading to widespread disappointment and renewed protests (Martin, 2021).

- **George Floyd**, a 46-year-old Black man, died on May 25, 2020 in Minneapolis, Minnesota after a white police officer knelt on his neck for over nine minutes during an arrest for allegedly using a counterfeit $20 bill at a convenience store. Despite Floyd repeatedly stating that he couldn't breathe and pleading for his life, the officer continued to apply pressure to his neck, ultimately causing Floyd's death. Video footage of the incident captured by bystanders quickly went viral, leading to widespread condemnation and demands for justice. Protesters took to the streets in unprecedented numbers in cities across the United States and around the world, calling for an end to police violence against Black people and systemic racism in law enforcement. The officer was arrested and charged with second-degree murder, third-degree murder, and second-degree manslaughter. The three other officers involved in Floyd's arrest were also charged with aiding and abetting second-degree murder and manslaughter. On April 20, 2021, after deliberating for approximately ten hours, the jury found the officers were found guilty on all charges. The verdict marked a rare instance of a police officer being held criminally accountable for the death of a Black person in the United States (Cheung, 2020; McGreal et al., 2021).

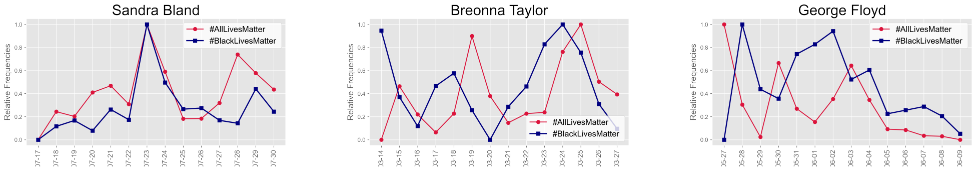


Figure 3.3: Normalized counts for the three events of interest.

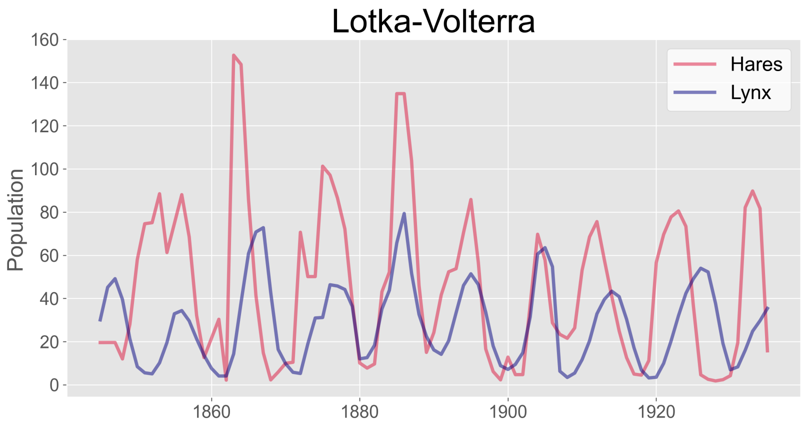


Figure 3.4: Lynx and hare counts over time from 1801 to 1930, which inspired the creation of the Lotka-Volterra model.

3.3.2 Methods

The Lotka-Volterra equations, also known as the predator-prey model, are a pair of first-order, non-linear, differential equations frequently used to describe the dynamics of biological systems involving two species in competition or interaction with each other. The model was independently developed by Alfred J. Lotka and Vito Volterra in the early 20th century. At its core, the Lotka-Volterra model typically involves two variables representing the populations of two interacting species over time. One variable represents the population of the “predator" species (e.g., lynx), while the other represents the population of the “prey" species (e.g., hares). The model describes how changes in the populations of these species affect each other over time (Wangersky, 1978).

The basic form of the Lotka-Volterra model includes parameters that represent factors such as the reproduction rate of the prey, the rate at which predators consume prey, and the death rate of predators in the absence of prey. These parameters influence the growth or decline of each species’ population, creating dynamic oscillations in the predator and prey populations. The model assumes certain simplifications, such as constant parameters and populations, which may not fully capture the complexities of real ecosystems. The basic form of the equations is given by:

$$\begin{aligned}\frac{dH}{dt} &= \alpha H - \beta HL, \\ \frac{dL}{dt} &= \delta LH - \gamma L,\end{aligned}$$

where:

- $H(t)$ represents the population of prey (e.g., hares),
- $L(t)$ represents the population of predators (e.g., lynx),
- α, β, δ, and γ are positive constants representing the rates of prey reproduction, predation, predator reproduction, and predator mortality, respectively.

With initial conditions:

$$\begin{aligned}H(0) &= H_0, \\ L(0) &= L_0,\end{aligned}$$

where H_0 and L_0 are the initial populations of “prey" and “predators", respectively.

We use the Lotka-Volterra models to represent the dynamics of #BlackLivesMatter and #AllLivesMatter over time. To do so, we perform a parameter estimation problem and minimize the sum of squared errors using the following equation sum of squared error (SSE) between true data y_i and parameter estimation data $\hat{y}_i$ using:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

We obtain counts for uses of #BlackLivesMatter and #AllLivesMatter and treat them as populations, with #BlackLivesMatter acting as the “predator" and #AllLivesMatter acting as the “prey". To obtain such counts, we utilize the Twitter API and obtain counts at a granularity of daily. Due to the fluctuation of values by orders of magnitude, we begin by normalizing our data. Given a dataset C, we obtain the normalized value c_N from a value $c \in C$, by utilizing the following formula:

$$c_N = \frac{\max(C) - \min(C)}{c - \min(C)}.$$

This formula scales each value in the dataset between 0 and 1.

To obtain an initial guess for parameters α, β, δ, and γ, we first truncate the data to include only the middle values (omitting the first and last few days). For each event, we choose H_0 and L_0 by observing when online chatter takes off (i.e. a significant quantity of hashtag use occurs). We find that this is often around 0-5 days after the actual event. We then compute the period of the hare populations T_H, and divide by 2π to obtain the frequency, ω.

We then locate the two lowest “predator" populations and obtain corresponding “prey" populations to solve

$$H(0) = H_0 e^{\alpha t}.$$

Using this value of α, we obtain the remaining values as such:

$$\begin{aligned}\beta &= \frac{\alpha}{\mu_L}\\ \delta &= \frac{\omega^2}{\alpha}\\ \gamma &= \frac{\delta}{\mu_H}.\end{aligned}$$

Utilizing this informed initial guess, we use `fminsearch` to perform the parameter estimation problem.

3.3.3 Results

Each of the models converged. Parameters α, β, δ, and γ should be *positive* numbers presenting predator-prey dynamics; however, we observe that is the case only for George Floyd. From this, we can infer that in the cases of Sandra Bland and Breonna Taylor, #AllLivesMatter tweet counts, act, in fact, as the “predator" population. Consequently, #BlackLivesMatter tweet counts act as the prey.

Table 3.2: Converged function values from the parameter estimation problem.

	α	β	δ	γ
Sandra Bland	-0.5869	-2.2503	-3.8517	-9.2795
Breonna Taylor	-1.0437	-2.1652	-1.5417	-3.9588
George Floyd	2.1940	5.3960	1.6157	5.4785

When predators interact with prey, this results in a decline in the prey population and an increase in the predator population. We find differences in which hashtag truly acts as the “predator" and which acts as the “prey". For the case of George Floyd, the #BlackLivesMatter conversation act to drown out or dwindle the #AllLivesMatter conversation. Conversely, for the cases of Sandra Bland and Breonna Taylor, the #AllLivesMatter conversation acts to dictate overall discourse and detract attention from #BlackLivesMatter discourse.

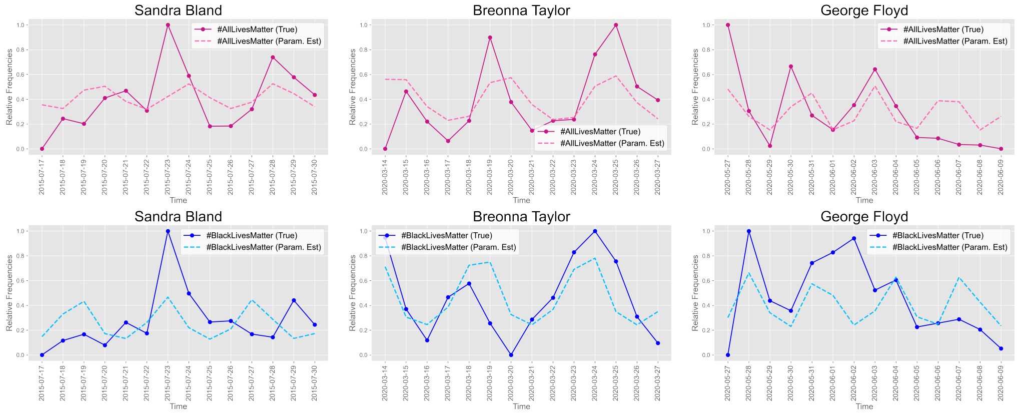


Figure 3.5: 14-day window of true data and parameter estimation data for each event.

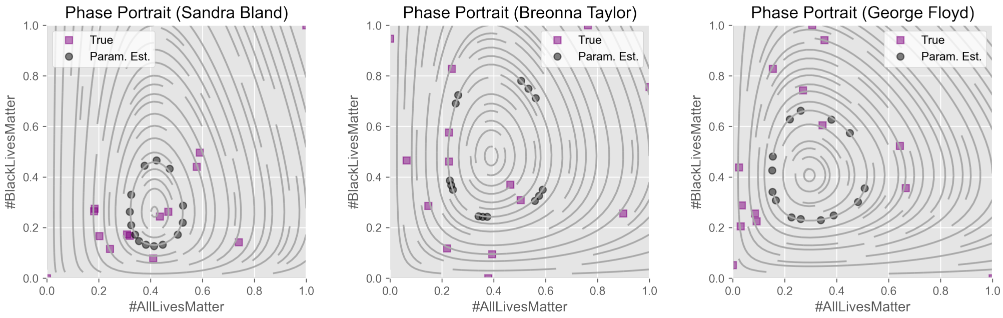


Figure 3.6: Streamplots for each event, with true data and parameter estimation data shown. Streamplots serve as a graphical tool to illustrate the dynamics of interactions between these populations over time. In our case, the trajectories visualize the flow of ideas conversation between different online communities.

Moreover, the Lotka-Volterra models assume uniform fluctuation, which isn't consistently the case for true tweet volume values. We find that parameter values tend to fit the mean values over time, rather than capturing fluctuations in periods.

3.4 Hashtags as Signals of Political Identity

3.4.1 Motivation

In the contemporary United States, political partisanship has become one of the most salient identity categories, correlating with variation on traits from religiosity to gun ownership to television show preference (DellaPosta, Shi, & Macy, 2015; Mason, 2018; Blakley et al., 2019). Accordingly, Americans on the political left and right appear to inhabit very different mental worlds. Differences in psychological traits, including need for cognition, tolerance for ambiguity, and need to evaluate, have been found to correlate with differences in political ideology (Young et al., 2019). Further, left-right political orientation appears to correlate with reliably different personality profiles, resulting in correspondingly different behavioral patterns (Carney et al., 2008). The phenomenon of affective polarization is at this point well described, whereby political decisions of left and right partisans are driven more by opposition to the other side than by any positive policy preferences (Abramowitz & Webster, 2016; Iyengar et al., 2019; Osmundsen et al., 2021). Moreover, identical stimuli can be perceived in a dramatically different light by left and right partisans (Van Bavel & Pereira, 2018). For example, Kahan *et al.* (Kahan et al., 2012) presented participants with identical footage of a protest and asked about their support for police intervention to quell it. Republican participants were more likely than Democrats participants to support police action when told the protest was in opposition to the military's policy of "don't ask, don't tell" outside a military recruitment office, while the effect was reversed when participants were told that the protesters were opposing abortion outside an abortion clinic.

Although neither #AllLivesMatter nor #BlackLivesMatter as movements are formally associated with any political party, they have over time become entangled in the increasingly polarized landscape of American political identity (Gallagher et al., 2018; S. Kim & Lee, 2021). Recent studies found that Democrats show increased support for the Black Lives Matter movement compared with Republicans (Drakulich & Denver, 2022; Thomas & Horowitz, n.d.), though neither study looked specifically at hashtags. Less evidence exists about partisanship and the All Lives Matter movement, though a recent qualitative analysis argued that the movement has been far more often invoked by Republican political candidates than by Democrats (Paul, 2019). Given the extent of polarization in the U.S. around political identities, it seems possible not only that perceptions of the two hashtags may differ wildly between left and right partisans, but even that the hashtags themselves may serve as a sort of identity signal, providing reliable context cues regarding how the author of an online message wishes their statement to be interpreted. Because this form of communications on Twitter surrounding these hashtags can be emotive and are so commonplace, studying its characteristics and evolution can be incredibly insightful into the way in which individuals navigate around online disagreement about sensitive topics.

We report on our investigations into how political identity moderates the perception of tweets tagged with the #BlackLivesMatter and #AllLivesMatter hashtags, expecting that partisans on the left would view the former more favorably than the latter, with the reverse effect for partisans on the political right. We were particularly interested in participants' perceptions of the tweets as offensive or racist. Moreover, we investigated the specific information content of the hashtags themselves in fueling partisan perceptions. We did this by artificially removing the hashtags from tweets in which they initially appeared, as well as by appending them to tweets completely unconnected to either movements. We investigated a number of possible predictors of affective responses to tweets, with a particular emphasis on political identity—an emphasis that, as we shall see, appears to have been warranted.

3.4.2 Methods

Dataset

To obtain a dataset of #AllLivesMatter and #BlackLivesMatter tweets, we used a web crawler(“jonbakerfish/TweetScraper: TweetScraper is a simple crawler/spider for Twitter Search without using API”, n.d.), which obtains only publicly available tweets via Twitter Advanced Search in compliance with Twitter’s rules[2]. We focused on tweets published in the year 2020 in order to constrain the contextual meaning of the tweets to be maximally salient to our participants, who evaluated the tweets in early 2021. That is, we scraped tweets containing either hashtag (“#AllLivesMatter" or “#BlackLivesMatter", case insensitive), and published between January and December 2020. This resulted in a total of 24 queries (one for each month for each hashtag) and yielded a total of 3,515,489 tweets (2,963,778 #BlackLivesMatter tweets and 551,711 #AllLivesMatter tweets). We then filtered these to create a set of tweets that contained only one hashtag, and had no mention of other Twitter handles and no attachments (pictures, videos, links, etc.). We further filtered the set of tweets manually, so that all tweets placed the hashtag at the very end of the tweet and did not use the hashtag itself as the subject of the tweet’s message (*e.g.*, “My least favorite hashtag is #BlackLivesMatter"). In other words, our interest was in tweets that used the hashtags only as concluding tags.

Neutral tweets were sampled from previous studies in which tweets were evaluated via crowdsourcing and rated as being racist, sexist, both, or neither (Waseem, 2016; Waseem & Hovy, 2016). We selected tweets from these datasets that were not rated by any participant as either racist or sexist and that appeared to us to be about politically neutral content. Some examples of the topics addressed in these tweets include the weather, food, and traffic.

We applied a sentiment analysis to the three groups of tweets (#AllLivesMatter tweets, #BlackLivesMatter tweets, and the set of neutral tweets), from the `nltk` package on Python, which utilizes `vader` to employ a word-lookup based scoring (Bird, Klein, & Loper, 2009). The results of that analysis are shown in Figure 3.7. We observe that all sets of tweets are generally more negative than positive in sentiment. Additionally, we observe minimal differences between sentiment distributions of #AllLivesMatter and #BlackLivesMatter tweets, diminishing the possibility that any differences in the interpretation of these tweets is due to differences in their overall sentiment. The distributions of positive and negative sentiment scores for #AllLivesMatter and #BlackLivesMatter tweets were more similar to one another than either were to the neutral tweets, which perhaps unsurprisingly tended to express substantially weaker sentiments overall.

Each set of tweets was further reduced to a small sample for use in participant surveys, for which we used 300 tweets in total. These were partitioned into ten distinct sets comprised of 30 tweets each. Each set contains 13 #AllLivesMatter tweets, 13 #BlackLivesMatter tweets,

[2] https://help.twitter.com/en/rules-and-policies/twitter-search-policies

and four neutral tweets. The size of these sets was based on the number of tweets our pilot study determined could be reasonably rated by participants without fatigue or attrition, in order for each tweet to be rated by multiple participants. Each participant was randomly assigned one of the ten distinct sets of tweets to evaluate, either with or without hashtags present.

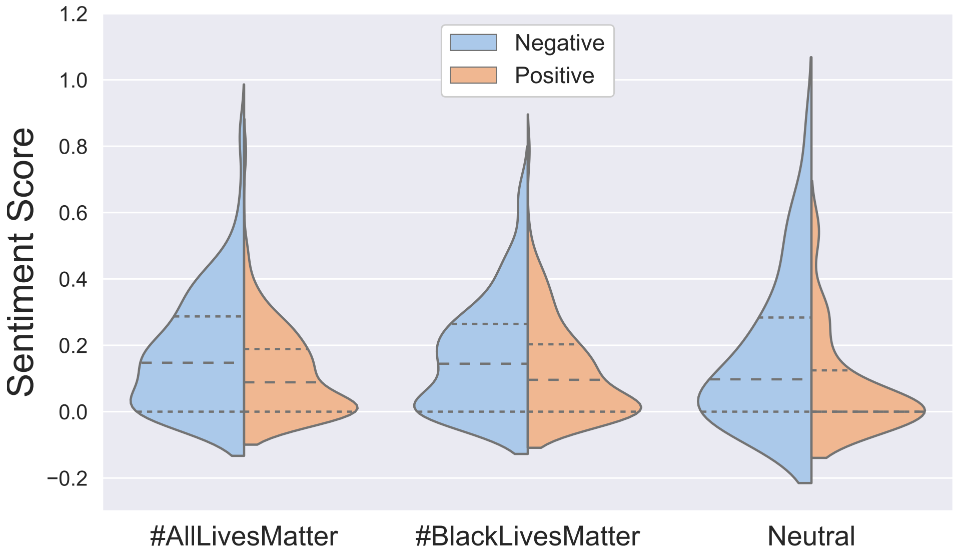


Figure 3.7: **Tweet sentiment scores.** Violin plots of positive and negative sentiment scores for #AllLivesMatter, #BlackLivesMatter, and neutral tweets used for this study. Dashed lines represent the means and dotted lines delineate the upper and lower quartiles of each distribution.

Survey Setup

At the beginning of the survey, each individual was asked to submit written consent to participate in the study. Individuals were prompted to select either “I consent to participate in this study" or “I do not wish to participate in this study" after being shown descriptions of the study’s purpose, procedures, compensation, risks, benefits, and confidentiality. They were also given the right to refuse or withdraw from the study. Following the consent portion, users were then prompted to complete a CAPTCHA verification. If the individual denied consent, the survey ended immediately. If the individual agreed and successfully completed verification, they were next provided with detailed instructions on how to complete the study, as well a necessary definitions. Participants were then presented with 30 tweets in random order and asked to evaluate them on several criteria. For each tweet, participants were instructed to evaluate whether its contents could be perceived as racist, offensive, both or neither, and whether these perceptions applied to (i) themselves personally, (ii) individuals within their social network, and (iii) individuals outside of their social network. The terms “personally", “within social network", and “outside of social network" were defined in the instructions. Our goal in asking participants to imagine how other people were likely to perceive the tweets was to enable us to examine the extent to which participants viewed their own valuations as being related to their social identities rather than as either solely personal views or human universals.

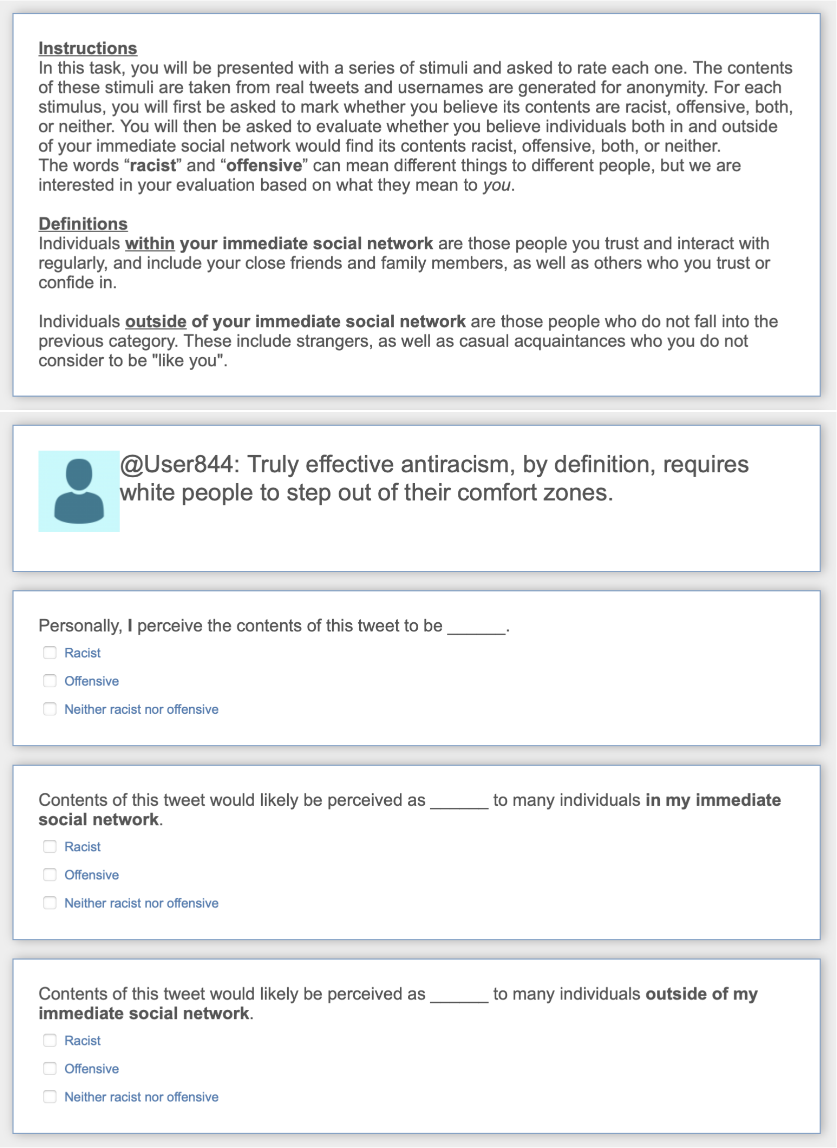
Participants were randomly assigned one of the ten datasets. To document the effect of hashtags on perceptions, some participants were presented tweets with hashtags and the others tweets without hashtags. If a participant was assigned the dataset with hashtags present, they were shown the raw tweets with hashtags already present and neutral tweets with "#AllLivesMatter" or "#BlackLiveMatter" appended. If a participant was assigned the dataset without hashtags present, they were shown the #AllLivesMatter and #BlackLivesMatter tweets with the hashtag omitted and unaltered neutral tweets.

After completion of tweet evaluations, participants were asked to fill out a demographic survey. Individuals were asked about their age, gender, familiarity with hashtags, news consumption, religiosity, and political orientation. We intentionally place the demographic survey *after* the tweet evaluations to ensure participants were not primed to give "identity-typical" responses.

To measure political orientation, participants were shown two opposing opinions (one "Conservative" take and one "Liberal" take) on 10 different political topics (see Table A.1) taken from a pre-existing PEW survey (Center, 2017). Then the participants were instructed with the following: "For each of the following, select the option that aligns most with your personal beliefs". Each participant started with a score of 0. For each Conservative opinion chosen, 1 was added to their score and for each Liberal opinion chosen, -1 was added to their score, resulting in a range of scores from -10 to 10 with -10 being as Liberal as possible and 10 being as Conservative as possible.

The distribution of these political orientation scores among this study's participants are shown in Fig. A.5. This distribution shows that the participants leaned Liberal with respect to this measure ($\mu = -3.966$). Nonetheless, there are a reasonable number of participants across this political orientation spectrum to study any behavioral trends with respect to political orientation score.

In the United States especially, religiosity tends to have significant, yet complex, effect on an individual's political views and general identity (Egan, 2020; Campbell et al., 2018; Lee et al., 2018). To gauge religiosity in a more fine-grained way, we utilized a subset of the Centrality of Religiosity Scale (CRS) (Huber & Huber, 2012), a measure of the importance of religion in a person's life. In order to focus on identity-relevant aspects, we selected questions that gauged participation in religious services and membership in religious communities and omitted questions about self evaluations of spirituality. To measure political orientation, we adapted an 11-question survey from the Pew Research Center (Center, 2017). Participants were shown a series of two opposing opinions (one "Conservative" take and one "Liberal" take) on 10 different political topics, and asked to select the option that best aligned with their personal beliefs. Each participant started with a score of 0. For each Conservative opinion chosen, 1 was added to their score



Instructions
In this task, you will be presented with a series of stimuli and asked to rate each one. The contents of these stimuli are taken from real tweets and usernames are generated for anonymity. For each stimulus, you will first be asked to mark whether you believe its contents are racist, offensive, both, or neither. You will then be asked to evaluate whether you believe individuals both in and outside of your immediate social network would find its contents racist, offensive, both, or neither.
The words "**racist**" and "**offensive**" can mean different things to different people, but we are interested in your evaluation based on what they mean to *you*.

Definitions
Individuals **within your immediate social network** are those people you trust and interact with regularly, and include your close friends and family members, as well as others who you trust or confide in.

Individuals **outside of your immediate social network** are those people who do not fall into the previous category. These include strangers, as well as casual acquaintances who you do not consider to be "like you".

@User844: Truly effective antiracism, by definition, requires white people to step out of their comfort zones.

Personally, **I** perceive the contents of this tweet to be ______.

□ Racist
□ Offensive
□ Neither racist nor offensive

Contents of this tweet would likely be perceived as ______ to many individuals **in my immediate social network**.

□ Racist
□ Offensive
□ Neither racist nor offensive

Contents of this tweet would likely be perceived as ______ to many individuals **outside of my immediate social network**.

□ Racist
□ Offensive
□ Neither racist nor offensive

Figure 3.8: Participant view of instructions (top) and one sample tweet evaluation (bottom) on Qualtrics.

and for each Liberal opinion chosen, -1 was added to their score, resulting in a range of scores from -10 to 10 with -10 being maximally Liberal and $+10$ being maximally Conservative. We considered participants to be Liberal if their score was less than 0 and Conservative if their score was greater than 0. A potential limitation of this survey is that it restricts political opinions to those promoted in mainstream media, and excludes more radical or outside views (Pew Research Center, 2021). Nevertheless, such scores capture a great deal of the variation in American political identity.

To measure the religiosity of each participant, we have used a subset of the Centrality of

Religiosity Scale (CRS) huber2012centrality, a measure of the centrality, importance or salience of religious meanings in personality. Table A.2 gives the selected questions for this survey.

Before distributing the survey, we recieved Institutional Review Board (IRB) approval from the UC Merced IRB (IRB#: UCM2020-70). We recruited a total of 1,428 participants through Amazon Mechanical Turk. All participants had to be located in the U.S., be over 18 years old, and have a HIT Approval Rate above 95%. We inserted two check questions into our survey to gauge a user's attentiveness to the survey in order to avoid users who randomly select choices without reading the survey content. If the individual got one or both question(s) wrong, we omitted their response. After performing omissions based upon check questions, a total of 1,244 viable participants remained. Our subsequent participant population was heavily skewed Liberal and White, while also being predominantly male.

3.4.3 Results

To understand the relationship between demographics and corresponding evaluations, we first examined the frequency of racist and offensive ratings as a function of individuals' demographic characteristics. Among all the demographic factors assessed, political orientation was the strongest predictor of whether tweets were perceived as racist or offensive. Perceptions of tweets marked with the #AllLivesMatter and #BlackLivesMatter hashtags were strongly mediated by political orientation, with individuals on the political left personally rating #AllLivesMatter tweets as being more offensive and racist than #BlackLivesMatter tweets. Conversely, individuals on the political right personally rated #BlackLivesMatter tweets as being more offensive and racist than #AllLivesMatter tweets. Results are shown in Figure 3.9.

When participants were asked to imagine how individuals within their personal social networks would respond to tweets, the patterns of ratings were nearly identical to their own personal evaluations, suggesting that our participants expect cohesion and agreement with those close to them (Figure 3.10, left). However, the association between political orientation and perceptions of tweets as racist or offensive did not hold when participants were asked to imagine how someone outside their social network would respond, suggesting individuals understood that their judgment of the tweets as racist or offensive would not be shared by everyone (Figure 3.10, right).

The effect of hashtag presence was most prevalent with left leaning participants when evaluating tweets marked with #AllLivesMatter as both racist and offensive (Figure 3.9, left column). A similar effect was observed with right leaning participants when evaluating tweets marked with #BlackLivesMatter as racist (Figure 3.9, bottom right). Overall, in both cases, the presence of the hashtag made the tweet contents more likely to be perceived as racist and/or offensive by partisans.

To verify that political orientation was the strongest predictor of how tweets were per-

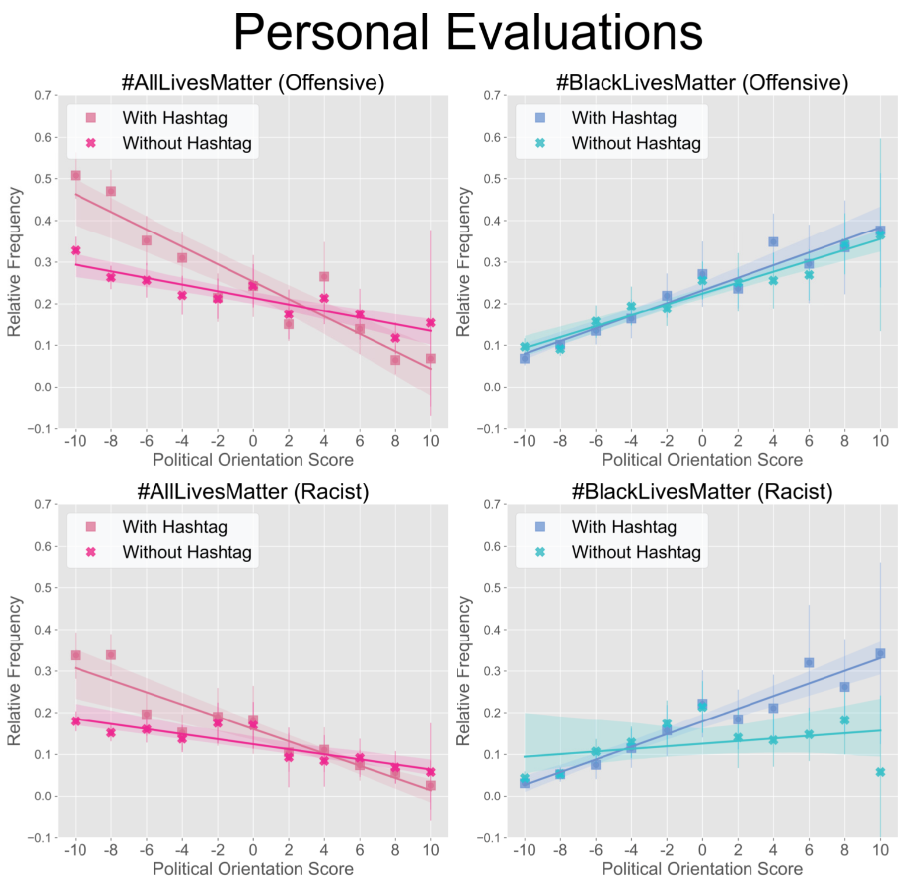


Figure 3.9: **Overall personal ratings by political orientation.** Relative frequencies of racist and offensive ratings for personal evaluations as a function of political score (with -10 being maximally Liberal and 10 being maximally Conservative), where relative frequency is calculated by dividing racist or offensive counts by total counts. 95% confidence is shown on relative frequencies and regressions.

ceived, we construct two sets of models: multivariate linear regression models and random forests models to predict racist and offensive evaluations as a function of age, gender, race, four different religiosity variables, and political orientation. For the multivariate linear regression models, we performed a partial f-test on all possible nested models (reduced models where one or more of the 8 demographic variables are removed). To evaluate the results, we examined both the f-statistic (a measure of error made by an individual nested model in terms of the residual sum of squares compared with the full model, where larger values are favorable) and p-value (a measure representing the probability that similar results would be observed if no effect was present, where smaller values are favorable). For each partial f-test, the largest (most favorable) f-statistic corresponded to the nested models that included all variables *except* political orientation score, which shows that the nested model with the most error compared with the full model does not consider

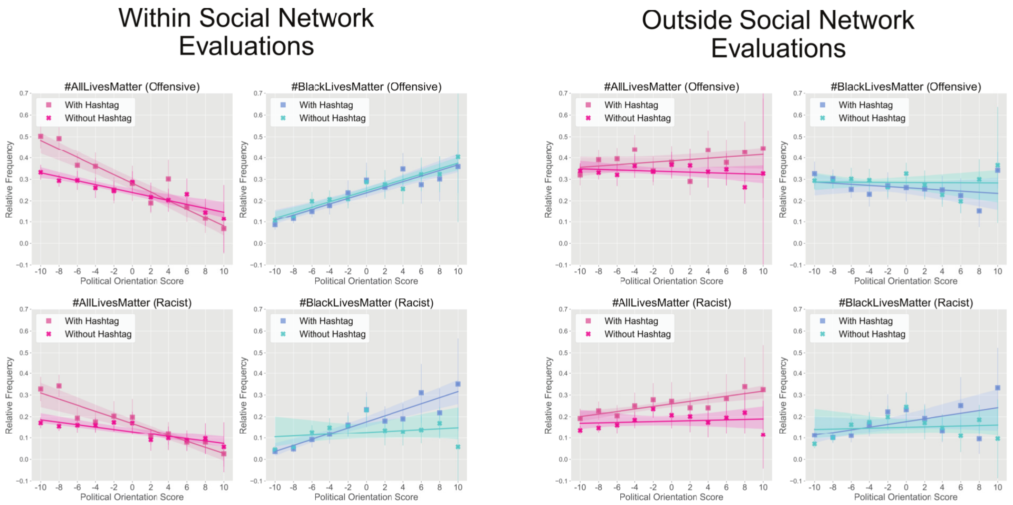


Figure 3.10: **Overall "within" and "outside" personal social network ratings by political orientation.** Relative frequencies of racist and offensive ratings for within personal social network and outside of personal social network evaluations as a function of political score (with -10 being maximally Liberal and 10 being maximally Conservative), where relative frequency is calculated by dividing racist or offensive counts by total counts. 95% confidence is shown on relative frequencies and regressions.

political orientation score. This indicated that political orientation score is the variable that has the strongest effect on participants' evaluations. Corresponding p-values were very small, with a maximum of 2.1629×10^{-6} and a minimum of 4.7837×10^{-27}. For random forest models, we evaluated feature importances. Results are shown in Figure 3.11, where each row corresponds to one model and gives the fractional amount of importance for each of the 8 feature or predictor variables, so that they sum to one. For each of the random forests models, the political orientation score is ranked substantially higher than all other predictors.

In Fig. A.5 we give the correlations between each demographic: age, gender, race, political score, and the four religiosity scores. Unsurprisingly, we found that the four religiosity scores to have the highest correlations to one another.

Given that political orientation yielded the strongest correlation to perceptions and that religiosity (beyond to itself) is correlated most to political orientation, we have studied the dependence on perception with religiosity.

In Fig. **??** we show perception results for #AllLivesMatter tweets with and without hashtags and #BlackLivesMatter tweets with and without hashtags. These results show that religious participants tended to perceive #BlackLivesMatter tweets racist and/or offensive, particularly for tweets with hashtag present, and were less likely to find #AllLivesMatter tweets racist and/or offensive. Conversely, non-religious participants are less likely to find #BlackLivesMatter racist and/or offensive, particularly for tweets with hashtag present, and were more likely to find #AllLivesMatter racist and/or offensive. We additionally note that the presence of a hashtag has more of an effect when evaluating #AllLivesMatter tweets than #BlackLivesMatter tweets.

We separated participants identifying as “white” from all others which we call “not white.” Figure A.4 shows perception results of #AllLivesMatter tweets with and without hashtags and #BlackLivesMatter tweets with and without hashtags. These results show that white participants were more likely to find #BlackLivesMatter racist and/or offensive and less likely to find #AllLivesMatter racist and/or offensive. Conversely, non-white participants were less likely to find #BlackLivesMatter racist and/or offensive and more likely to find #AllLivesMatter racist and/or offensive. The presence of hashtag has more of an effect when evaluating #AllLivesMatter tweets than #BlackLivesMatter tweets.

Both the multivariate linear regression models and the random forests models evince that political orientation is the strongest predictor amongst measured demographics of tweet evaluations. To analyze reliability of the tweet evaluations made by the participants, we compute intraclass correlation coefficients (ICC) (**koo2016guideline**; **liljequist2019intraclass**) among groups of individuals who were randomly assigned and thus evaluated the same set of tweets. The ICC is a statistical value between 0 and 1 that measures consistency of evaluations across multiple participants, with a measure of 0 indicating results are completely unreliable and a measure of 1 indicating perfect reliability. We compute ICC values from two models: a two-way random model (ICC$(2, k)$) and a two-way mixed model (ICC$(3, k)$), which differ based upon whether the groups of k participants are regarded as being representative of the entire population or as being the only participants of interest, respectively. In both cases, we find uniformly high values (> 0.90) across datasets for both racist and offensive ratings, strongly indicating that these tweet evaluations are reliable. Moreover, correlations between responses and other demographics (age, gender, etc.) either did not emerge in these analyses or were not significant in both of these models.

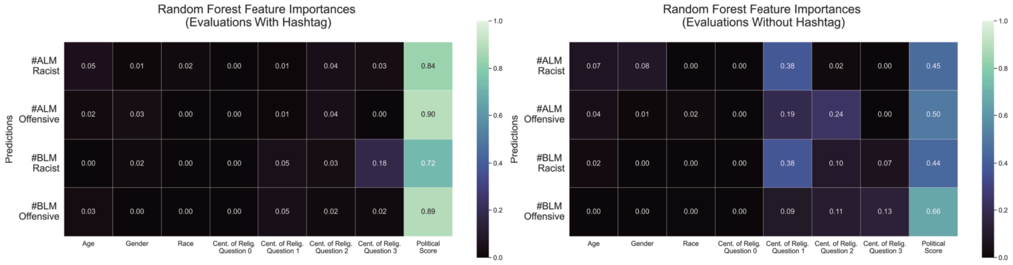
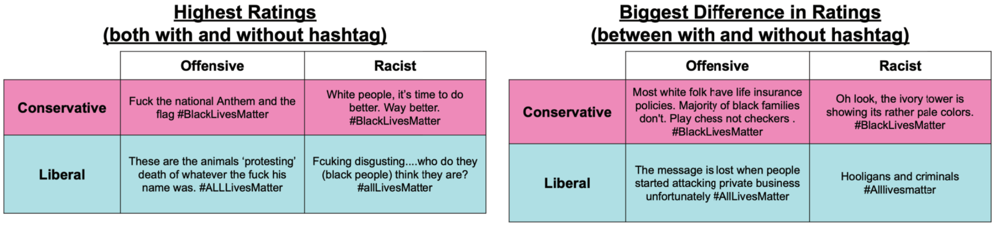


Figure 3.11: **Random Forest feature importances.** Results of feature importances for full random forests models (using all eight predictors). Feature importances lie between [0,1] and sum to 1, where 0 indicates a feature is not important at all and 1 indicates that a feature is as important as possible.

The text from some of the tweets used in our study can be viewed in Figure 3.12. The left side of this figure shows the tweets that were consistently rated as the most offensive or racist by right and left partisans. The right side of the figure shows the tweets that exhibited the largest differences in ratings between the hashtag and no-hashtag conditions. These tweets highlight that the information content of the hashtag can vary considerably. In some cases, a hashtag simply reinforces an already-clear message, while in other cases it contextualizes and clarifies an otherwise-ambiguous message.



Highest Ratings
(both with and without hashtag)

	Offensive	Racist
Conservative	Fuck the national Anthem and the flag #BlackLivesMatter	White people, it's time to do better. Way better. #BlackLivesMatter
Liberal	These are the animals 'protesting' death of whatever the fuck his name was. #ALLLivesMatter	Fcuking disgusting....who do they (black people) think they are? #allLivesMatter

Biggest Difference in Ratings
(between with and without hashtag)

	Offensive	Racist
Conservative	Most white folk have life insurance policies. Majority of black families don't. Play chess not checkers . #BlackLivesMatter	Oh look, the ivory tower is showing its rather pale colors. #BlackLivesMatter
Liberal	The message is lost when people started attacking private business unfortunately #AllLivesMatter	Hooligans and criminals #Alllivesmatter

Figure 3.12: **Significant tweet ratings.** On left, individual tweets with the highest frequency of offensive or racist ratings, regardless of hashtag presence (relative frequencies of > 0.9, > 0.76, > 0.86, > 0.84, respectively). On right, individual tweets for which hashtag presence made the largest difference in rating frequencies (differences in relative frequencies of 0.615, 0.488, 0.426, 0.412, respectively).

Unsurprisingly, neutral tweets were much less likely to be rated as racist or offensive than #AllLivesMatter and #BlackLivesMatter tweets (Figure 3.13). However, when one of these hashtags was artificially added to a neutral tweet, that tweet was more likely to be evaluated as racist or offensive. In particular, the addition of "#AllLivesMatter" to neutral tweets was associated with a large increase in ratings as racist or offensive among Liberal participants, while the addition of "#BlackLivesMatter" to neutral tweets was associated with a moderate increase in ratings of racist and offensive among both Liberal and Conservative participants. While we found it surprising that the addition of "#BlackLivesMatter" would increase perceptions of neutral tweets as racist and offensive among Liberal participants, it is possible that such responses are provoked by the juxtaposition of something deemed quite serious (the hashtag) in a banal context.

3.4.4 Discussion

In the United States and elsewhere, particularly in otherwise diverse nations, political identity is increasingly the dominant identity driving much of social behavior (Mason, 2018; Abramowitz & Webster, 2016; Iyengar et al., 2019; Joireman, 2003). Here, we have shown that among U.S. participants, perceptions of race-relevant hashtags #BlackLivesMatter and #AllLivesMatter diverge considerably in ways that are predicted by political orientation. Tweets

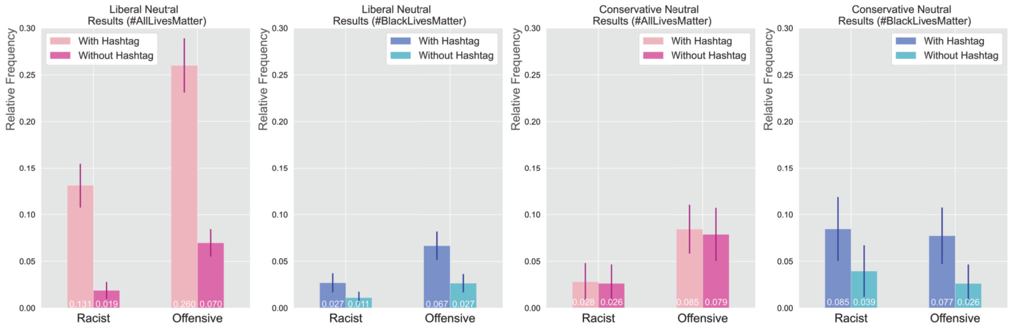


Figure 3.13: **Neutral tweet evaluations.** Evaluations of neutral tweets by political score with 90% confidence interval, where relative frequency is calculated by dividing racist or offensive counts by total counts. Independent t-tests revealed that there were statistically significant differences between Liberal participants evaluating neutral tweets with hashtags appended versus without, with the addition of "#AllLivesMatter" having a more significant effect (corresponding p-values of 2.359×10^{-13} and 1.746×10^{-21} for racist and offensive evaluations, respectively) than the addition of "#BlackLivesMatter" (corresponding p-values of 0.027 and 0.0002 for racist and offensive evaluations, respectively). Differences between Conservative participants evaluating neutral tweets with hashtags appended versus without were much less significant, with the addition of "#AllLivesMatter" having the weakest effect (corresponding p-values of 0.915 and 0.812 for racist and offensive evaluations, respectively), followed by the the addition of "#BlackLivesMatter" (corresponding p-values of 0.102 and 0.028 for racist and offensive evaluations, respectively).

tagged with #BlackLivesMatter were more likely to be rated as offensive and racist by participants on the political right, while tweets tagged with #AllLivesMatter were more likely to be rated as offensive and racist by participants on the political left. Political orientation was more strongly predictive of these divergent responses than any other demographic factors we examined, including the age, gender, religiosity, or race of the participants. Moreover, our results suggest that these trends are likely to be driven by identity-based assessments rather than more general perceptual differences between right and left partisans, because our main effect held when people were asked to imagine how someone else in their social networks would respond to the tweets, but not when they imagined how someone outside their social networks would response. Although other identity categories, notably historically persecuted identities associated with race and sexual orientation, are also associated with perceptions of the BLM and ALM movements in both Black and White participants (BONILLA & TILLERY, 2020; Cole, 2020; West, Greenland, & van Laar, 2021), political affiliation remains the strongest predictor of that support (West, Greenland, & van Laar, 2021).

The associations between political orientation and the tweet ratings were severely, though not entirely, diminished when the hashtags themselves were removed from the text of the tweets. However, the effect of hashtag was not consistent from tweet to tweet. In some cases, hashtags serve merely to amplify an already-clear meaning, while also increasing searchability. In other cases, however, the meaning of a tweet was ambiguous in the absence of the hashtag. In these cases, a hashtag serves to contextualize the tweet's text and suggest a particular race-related interpretation. This role appears to have been especially important for tweets where the ratings between the hashtag and no-hashtag conditions were very different. So, although tweets marked with #BlackLivesMatter and #AllLivesMatter hashtags had stronger negative valences than neutral tweets, responses to tweets marked with these hashtags were not merely driven by the text communicated in those tweets. The hashtags themselves served as important signals, as indicated both by the diminishment of the main effect when hashtags were removed from the original tweets as well as the reintroduction of the effect when hashtags were added to neutral tweets.

Both #BlackLivesMatter and #AllLivesMatter are ostensibly about race, so it is perhaps unsurprising that removal of either hashtag reduced ratings of tweets as racist by right and left partisans, respectively. While the presence of the #BlackLivesMatter hashtags was also predictive of ratings of tweets as offensive by right partisans, these ratings appear to be driven largely by the content of the tweets themselves, and not by the hashtag. This was not the case for #AllLivesMatter, the presence of which was associated with a large increase in left partisans' ratings of a tweet as offensive. Individuals on the political left appear to have a particularly strong reaction to the #AllLivesMatter hashtag, finding its presence offensive even when it is attached to otherwise neutral tweets. This indicates that among left partisans, #AllLivesMatter is seen not only as a marker that contextualizes other communication, but as an offensive statement in its own right. Partisans on the right may find the #BlackLivesMatter hashtag racist because they believe there is an implicit “only" in front of “black lives matter," while left partisans may be more likely to tacitly append the statement with “too."

The suite of views associated with political identity is not stable and particular signals are not likely to be associated with any given identity forever. Our study, however, does illuminate an association between identity, viewpoints, and signals at this point in time, which can inform our understanding of politically-relevant communication both on- and offline. More generally, our study helps to demonstrate the extent to which identity—including political identity within an allegedly integrated society—can dramatically shape how information is processed and interpreted. This can have important societal ramifications, as rational conversations about important concepts require firm grounding in how individuals are using particular terms. For example, when asked to name “socialist" countries, the top three answers given by Republican voters in the U.S. were Venezuela, China, and Russia, while the top three answers given by Democratic voters were Denmark, Sweden, and Norway (Matthew Smith, 2020). Such divergent usage of the same word limits the ability of Americans to engage in meaningful dialogue about the pros and cons of socialist policies. Similarly, disagreements about what is meant by “Black Lives Matter" or “All Lives Matter", as well as what is or is not racist or offensive is likely to hinder the ability of Americans to reach consensus or even compromise on these and related issues.

Chapter 4

Conclusion

4.1 Summary

Amidst the vast expanse of online content, individuals increasingly turn to the internet as a source of information, guidance, and community. Whether seeking answers to pressing questions, exploring diverse perspectives, or connecting with like-minded individuals, the digital realm offers unparalleled access to knowledge and resources. This reliance on online platforms reflects not only a shift in information consumption patterns but also underscores the evolving role of technology in shaping human behavior and societal norms. In the digital age, social media platforms have emerged as powerful catalysts for social movements, revolutionizing the landscape of activism and advocacy. The instantaneous connectivity afforded by platforms like Twitter, Facebook, and Instagram enables the rapid dissemination of information, mobilization of supporters, and amplification of marginalized voices. From grassroots campaigns to global protests, social media serves as a virtual agora, facilitating dialogue, organizing collective action, and galvanizing social change (Chon & Park, 2020; Kidd & McIntosh, 2016; Madison & Klang, 2020).

Demographics wield significant influence over our interactions with the world, shaping our perceptions, experiences, and opportunities. Factors such as age, gender, ethnicity, and socioeconomic status profoundly impact how individuals navigate societal structures and engage with their communities. Demographics not only shape our experiences but also influence how we perceive, interact with, and contribute to the world around us. Recognizing and understanding the diverse ways in which demographic factors affect and shape human experiences is essential for comprehending the complexities of human interaction and societal dynamics. This broader understanding helps in creating a more inclusive and equitable society for everyone.

4.2 Future Work

Over the past few years, the field of natural language processing (NLP) has undergone significant transformations, driven primarily by the development and adoption of advanced models such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). More recently, the rise of generative large language models, exemplified by models like ChatGPT and Gemini, marks a new frontier in NLP research (Wu et al., 2023; McIntosh et al., 2023). The progress achieved through these models offers unprecedented opportunities to enhance and refine NLP applications that were previously reliant on more rudimentary techniques. With the sophisticated understanding facilitated by transformer models and generative language models, we can now analyze and interpret text, conversations, and user interactions at a depth that was inconceivable just a few years ago. The enhanced capabilities of these models thus promise significant benefits for our social understanding of discourse and user behavior online. This allows for more sophisticated understandings and studying of online discourse, helping to identify trends, detect misinformation, and understand public sentiment more than ever before. The potential for innovation in understanding and improving social interactions and behaviors online grows heralds a new era of NLP research and development.

The exploration of future directions in understanding online behavior across various micro-blogging platforms is integral to anticipating the evolving dynamics of social media. Over recent years, the digital landscape has undergone significant transformations, marked notably by shifts such as the downfall of Twitter and the emergence of alternative platforms Threads and Bluesky (ORTUTAY, 2023; Mehta, 2023). This diversification challenges the once-centralized influence of a single platform like Twitter and prompts reflection on the future trajectory of online interactions. With no singular micro-blogging platform asserting universal dominance, we must reflect on the emerging contours of this evolving ecosystem and how our present findings may inform future inquiries. Despite this dispersion, the underlying principles governing user behavior and linguistic patterns remain pertinent. Understanding these fundamental aspects not only provides insight into current social phenomena but also serves as a foundation for extrapolating future trends.

As long as oppression exists, there will be a continuous and compelling need to study social movements. This reality ensures that there is a continual influx of rich data for scholars and activists committed to studying and combating these injustices. The diverse manifestations of oppression provide a fertile ground for research into the mechanisms of resistance and social change. Movements offer invaluable insights into the dynamics of collective action and the outcomes of sustained activism. Historical and ongoing struggles present an ever-expanding repository of case studies. Each social movement, with its unique context and challenges, enriches the understanding of how people organize, resist, and strive for a more equitable world. Moreover, the digital age has amplified both the visibility of oppression and the capacity for mobilization. Social media platforms, for instance, have become crucial tools for documenting injustices, spreading awareness, and coordinating activism on a global scale (Yılmaz, 2017).

The impact of online dynamics extends beyond the digital realm, influencing real-world events and social movements. Exploring the intersection of online discourse and offline consequences sheds light on the broader societal implications of digital interactions. Understanding the mechanisms underlying online behavior equips us to anticipate future trends and harness the potential of digital platforms for positive social change. By contextualizing our insights within broader socio-political dynamics, we can contribute meaningfully to discourse surrounding online engagement and its real-world ramifications.

Bibliography

Abramowitz, A. I., & Webster, S. (2016). The rise of negative partisanship and the nationalization of us elections in the 21st century. *Electoral Studies*, *41*, 12–22.

Adamska, K. (2015). Hashtag as a message? the role and functions of hashtags on twitter. *Media Studies/Studia Medioznawcze*, *62*(3).

Aggarwal, K., Singh, S. K., Chopra, M., & Kumar, S. (2022). Role of social media in the covid-19 pandemic: A literature review. *Data mining approaches for big data and sentiment analysis in social media*, 91–115.

Ahmed, W., Seguí, F. L., Vidal-Alaball, J., & Katz, M. S. (2020). Covid-19 and the "film your hospital" conspiracy theory: Social network analysis of twitter data. *Journal of medical Internet research*, *22*(10), e22374.

Ahmed, W., Vidal-Alaball, J., Downing, J., Seguí, F. L., et al. (2020). Covid-19 and the 5g conspiracy theory: Social network analysis of twitter data. *Journal of medical internet research*, *22*(5), e19458.

Aımeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, *13*(1), 30.

Ajibade, I., & Boateng, G. O. (2021). Predicting why people engage in pro-sustainable behaviors in portland oregon: The role of environmental self-identity, personal norm, and socio-demographics. *Journal of Environmental Management*, *289*, 112538.

Akbulaev, N., Mammadov, I., & Aliyev, V. (2020). Economic impact of covid-19. *Sylwan*, *164*(5).

Allington, D., Duffy, B., Wessely, S., Dhavan, N., & Rubin, J. (2021). Health-protective behaviour, social media usage and conspiracy belief during the covid-19 public health emergency. *Psychological medicine*, *51*(10), 1763–1769.

Anderson, R. M., Vegvari, C., Truscott, J., & Collyer, B. S. (2020). Challenges in creating herd immunity to sars-cov-2 infection by mass vaccination. *The Lancet*, *396*(10263), 1614–1616.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on twitter. *Proceedings of the fourth ACM international conference on Web search and data mining*, 65–74.

Balakrishnan, V., Abdul Rahman, L. H., Tan, J. K., & Lee, Y. S. (2023). Covid-19 fake news among the general population: Motives, sociodemographic, attitude/behavior and impacts –a systematic review. *Online Information Review*, *47*(5), 944–973.

Barth, F. (1969). *Ethnic groups and boundaries*. Little, Brown.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the international AAAI conference on web and social media*, *3*(1), 361–362.

Béland, L.-P., Brodeur, A., & Wright, T. (2023). The short-term economic consequences of covid-19: Exposure to disease, remote work and government response. *Plos one*, *18*(3), e0270341.

Benballa, M., Collet, S., & Picot-Clemente, R. (2019). Saagie at semeval-2019 task 5: From universal text embeddings and classical features to domain-specific text classification. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 469–475.

Berger, J., & Heath, C. (2008). Who drives divergence? identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *Journal of Personality and Social Psychology*, *95*(3), 593–607.

Bhat, P., & Klein, O. (2020). Covert hate speech: White nationalists and dog whistle communication on twitter. In G. Bouvier & J. E. Rosenbaum (Eds.), *Twitter, the public sphere, and the chaos of online deliberation* (pp. 151–172). Springer International Publishing. https://doi.org/10.1007/978-3-030-41421-4_7

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Blakley, J., Watson-Currie, E., Shin, H., Valenti, L., Saucier, C., & Boisvert, H. (2019). Are you what you watch? tracking the political divide through tv preferences. *Normal Lear Center*. https://learcenter.org/wp-content/uploads/2019/05/are%5C_you%5C_what%5C_you%5C_watch.pdf

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

BONILLA, T., & TILLERY, A. B. (2020). Which identity frames boost support for and mobilization in the blacklivesmatter movement? an experimental test. *American Political Science Review*, *114*(4), 947–962. https://doi.org/10.1017/S0003055420000544

Bowleg, L. (2020). We're not all in this together: On covid-19, intersectionality, and structural inequality.

Boyle, K., & Rathnayake, C. (2020). # Himtoo and the networking of misogyny in the age of# metoo. *Feminist Media Studies*, *20*(8), 1259–1277.

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

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., & Zhilin, O. (2020). The causes and consequences of covid-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*, *1*(3).

Bruder, M., Haffke, P., Neave, N., Nouripanah, N., & Imhoff, R. (2013). Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy mentality questionnaire. *Frontiers in psychology*, *4*, 225.

Burkolter, D., & Kluge, A. (2011). Online consumer behavior and its relationship with socio¬ demographics, shopping orientations, need for emotion, and fashion leadership. *Journal of Business and Media Psychology*, *2*, 20–28.

Campbell, D. E., Layman, G. C., Green, J. C., & Sumaktoyo, N. G. (2018). Putting politics first: The impact of politics on american religious and secular orientations. *American Journal of Political Science*, *62*(3), 551–565.

Carney, D. R., Jost, J. T., Gosling, S. D., & Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political psychology*, *29*(6), 807–840.

Carrasco-Farré, C. (2022). The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications*, *9*(1), 1–18.

Center, P. R. (2017). The partisan divide on political values grows even wider. *Pew Research Center*.

Charquero-Ballester, M., Walter, J. G., Nissen, I. A., & Bechmann, A. (2021). Different types of covid-19 misinformation have different emotional valence on twitter. *Big Data & Society*, *8*(2), 20539517211041279.

Chary, M. A., Overbeek, D. L., Papadimoulis, A., Sheroff, A., & Burns, M. M. (2021). Geospatial correlation between covid-19 health misinformation and poisoning with household cleaners in the greater boston area. *Clinical toxicology*, *59*(4), 320–325.

Cheung, H. (2020). George floyd death: Why us protests are so powerful this time. *BBC News*.

Chon, M.-G., & Park, H. (2020). Social media activism in the digital age: Testing an integrative model of activism on contentious issues. *Journalism & Mass Communication Quarterly*, *97*(1), 72–97.

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9).

Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W.-C., Wang, C.-B., & Bernardini, S. (2020). The covid-19 pandemic. *Critical reviews in clinical laboratory sciences*, *57*(6), 365–388.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

Cole, G. (2020). Types of white identification and attitudes about black lives matter. *Social Science Quarterly*, *101*(4), 1627–1633.

Cuan-Baltazar, J. Y., Muñoz-Perez, M. J., Robledo-Vega, C., Pérez-Zepeda, M. F., & Soto-Vega, E. (2020). Misinformation of covid-19 on the internet: Infodemiology study. *JMIR public health and surveillance*, *6*(2), e18444.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

DellaPosta, D., Shi, Y., & Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology*, *120*(5), 1473–1511.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Diseases, T. L. I. (2020). The covid-19 infodemic. *The Lancet. Infectious Diseases*, *20*(8), 875.

Dodd, M. D., Balzer, A., Jacobs, C. M., Gruszczynski, M. W., Smith, K. B., & Hibbing, J. R. (2012). The political left rolls with the good and the political right confronts the bad: Connecting physiology and cognition to preferences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1589), 640–649.

Donath, J. S. (1999). Identity and deception in the virtual community. In P. Kollock & M. Smith (Eds.), *Communities in cyberspace* (pp. 29–59). Routledge.

Dong, M., He, F., & Deng, Y. (2021). How to understand herd immunity in the context of covid-19. *Viral immunology*, *34*(3), 174–181.

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, *40*, 3–35.

Drakulich, K., & Denver, M. (2022). The partisans and the persuadables: Public views of black lives matter and the 2020 protests. *Perspectives on Politics*, 1–18.

Duckett, J., & Sacra, D. (2019). A cloud protest exploratory study examining the evolution of# takeaknee. *Pro Football and the Proliferation of Protest: Anthem Posture in a Divided America*, 175.

Duckitt, J., Bizumic, B., Krauss, S. W., & Heled, E. (2010). A tripartite approach to right-wing authoritarianism: The authoritarianism-conservatism-traditionalism model. *Political Psychology*, *31*(5), 685–715.

Egan, P. J. (2020). Identity as dependent variable: How americans shift their identities to align with their politics. *American Journal of Political Science*, *64*(3), 699–716.

Evanega, S., Lynas, M., Adams, J., Smolenyak, K., & Insights, C. G. (2020). Coronavirus misinformation: Quantifying sources and themes in the covid-19 'infodemic'. *JMIR Preprints*, *19*(10), 2020.

Face, H. (2024). Cardiffnlp/xlm-twitter-politics-sentiment [Accessed: 2024-06-17]. %5C%5C%20https://huggingface.co/cardiffnlp/xlm-twitter-politics-sentiment

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics reports*, *659*, 1–44.

Francis, M. M., & Wright-Rigueur, L. (2021). Black lives matter in historical perspective. *Annual Review of Law and Social Science*, *17*(1), 441–458. https://doi.org/10.1146/annurev-lawsocsci-122120-100052

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions.* WW Norton & Co.

Funkhouser, E. (2017). Beliefs as signals: A new function for belief. *Philosophical Psychology*, *30*(6), 809–831.

Gaitens, J., Condon, M., Fernandes, E., & McDiarmid, M. (2021). Covid-19 and essential workers: A narrative review of health outcomes and moral injury. *International Journal of Environmental Research and Public Health*, *18*(4), 1446.

Gallagher, R. J., Reagan, A. J., Danforth, C. M., & Dodds, P. S. (2018). Divergent discourse between protests and counter-protests:# blacklivesmatter and# alllivesmatter. *PloS one*, *13*(4), e0195644.

Garza, A. (2014). A herstory of the# blacklivesmatter movement.

Gauthier, G. R., Smith, J. A., García, C., Garcia, M. A., & Thomas, P. A. (2021). Exacerbating inequalities: Social networks, racial/ethnic disparities, and the covid-19 pandemic in the united states. *The Journals of Gerontology: Series B*, *76*(3), e88–e92.

Giansiracusa, N., & Giansiracusa, N. (2021). Social spread: Moderating misinformation on facebook and twitter. *How algorithms create and prevent fake news: Exploring the impacts of social media, deepfakes, GPT-3, and more*, 175–215.

Goertzel, T. (1994). Belief in conspiracy theories. *Political psychology*, 731–742.

Goffman, E. (1978). *The presentation of self in everyday life.* Harmondsworth.

Goswami, M. P. (2018). Social media and hashtag activism. *Liberty Dignity and Change in Journalism*, *2017*.

Gover, A. R., Harper, S. B., & Langton, L. (2020). Anti-asian hate crime during the covid-19 pandemic: Exploring the reproduction of inequality. *American journal of criminal justice*, *45*(4), 647–667.

Guasti, N. (2020). The plight of essential workers during the covid-19 pandemic. *Lancet*, (10237).

Hartman, R. O., Dieckmann, N. F., Sprenger, A. M., Stastny, B. J., & DeMarree, K. G. (2017). Modeling attitudes toward science: Development and validation of the credibility of science scale. *Basic and Applied Social Psychology*, *39*(6), 358–371.

Hillstrom, L. C. (2018). *The# metoo movement*. Bloomsbury Publishing USA.

Holbrook, C., Lopez-Rodriguez, L., & Gomez, A. (2018). Battle of wits: Militaristic conservatism and warfare cues enhance the perceived intellect of allies versus adversaries. *Soc. Psychol. Personal. Sci*, *8*, 670–678.

Huber, S., & Huber, O. W. (2012). The centrality of religiosity scale (crs). *Religions*, *3*(3), 710–724.

Iannaccone, L. R. (1992). Sacrifice and stigma: Reducing free-riding in cults, communes, and other collectives. *Journal of Political Economy*, *100*(2), 271–291.

Ince, J., Rojas, F., & Davis, C. A. (2017). The social media response to black lives matter: How twitter users interact with black lives matter through hashtag use. *Ethnic and racial studies*, *40*(11), 1814–1830.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, *22*, 129–146.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 56–65.

Jiang, J., Ren, X., Ferrara, E., et al. (2021). Social media polarization and echo chambers in the context of covid-19: Case study. *JMIRx med*, *2*(3), e29570.

Jin, D., Yu, Z., Jiao, P., Pan, S., He, D., Wu, J., Yu, P., & Zhang, W. (2021). A survey of community detection approaches: From statistical modeling to deep learning. *IEEE Transactions on Knowledge and Data Engineering*.

Joireman, S. (2003). *Nationalism and political identity*. A&C Black.

Jonbakerfish/tweetscraper: Tweetscraper is a simple crawler/spider for twitter search without using api. (n.d.).

Kahan, D. M., Hoffman, D. A., Braman, D., & Evans, D. (2012). They saw a protest: Cognitive illiberalism and the speech-conduct distinction. *Stanford Law Review*, *64*, 851.

Ki, H. K. (2021). Covid-19 vaccination and herd immunity. *The Journal of Korean Diabetes*, *22*(3), 179–184.

Kidd, D., & McIntosh, K. (2016). Social media and social movements. *Sociology Compass*, *10*(9), 785–794.

Kim, J. Y., & Kesari, A. (2021). Misinformation and hate speech: The case of anti-asian hate speech during the covid-19 pandemic. *Journal of Online Trust and Safety*, *1*(1).

Kim, S., & Lee, A. (2021). Black lives matter and its counter-movements on facebook. *Available at SSRN*.

Klein, G. C. (2018). On the death of sandra bland: A case of anger and indifference. *Sage open*, *8*(1), 2158244018754936.

Kobayashi, R., & Lambiotte, R. (2016). Tideh: Time-dependent hawkes process for predicting retweet dynamics. *Tenth International AAAI Conference on Web and Social Media.*

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, *120*(7), e2210666120.

Kuehn, B. M. (2020). Spike in poison control calls related to disinfectant exposures. *JAMA*, *323*(22), 2240–2240.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web*, 591–600.

Lamberty, P., & Leiser, D. (2019). Sometimes you just have to go in. conspiracy beliefs lower democratic participation and lead to political violence. In *Psyarxiv preprints*.

Lantz, B., & Wenger, M. R. (2023). Anti-asian xenophobia, hate crime victimization, and fear of victimization during the covid-19 pandemic. *Journal of interpersonal violence*, *38*(1-2), 1088–1116.

Lăzăroiu, G., Mihăilă, R., & Branişte, L. (2021). The language of covid-19 vaccine hesitancy and public health misinformation: Distrust, unwillingness, and uncertainty. *Review of Contemporary Philosophy*, *20*, 117–127.

Lee, K., Ashton, M. C., Griep, Y., & Edmonds, M. (2018). Personality, religion, and politics: An investigation in 33 countries. *European Journal of Personality*, *32*(2), 100–115. https://doi.org/https://doi.org/10.1002/per.2142

Lewandowsky, S., Gignac, G. E., & Oberauer, K. (2013). The role of conspiracist ideation and worldviews in predicting rejection of science. *PloS one*, *8*(10), e75637.

Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). Nasa faked the moon landing—therefore, (climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological science*, *24*(5), 622–633.

Lobato, E., Mendoza, J., Sims, V., & Chin, M. (2014). Examining the relationship between conspiracy theories, paranormal beliefs, and pseudoscience acceptance among a university population. *Applied Cognitive Psychology*, *28*(5), 617–625.

Lobato, E. J., & Zimmerman, C. (2019). Examining how people reason about controversial scientific topics. *Thinking & Reasoning*, *25*(2), 231–255.

Loury, G. C. (1994). Self-censorship in public discourse: A theory of “political correctness" and related phenomena. *Rationality and Society*, *6*(4), 428–461.

Mabrey, B. E. (2021). *The disinformation dozen and media misinformation on science and vaccinations* [Bachelor’s Thesis]. Oregon State University.

Madison, N., & Klang, M. (2020). The case for digital activism: Refuting the fallacies of slacktivism. *Journal of Digital Social Research*, *2*(2), 28–47.

Malliaros, F. D., & Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey. *Physics reports*, *533*(4), 95–142.

Martin, J. (2021). Breonna taylor: Transforming a hashtag into defunding the police. *J. Crim. L. & Criminology*, *111*, 995.

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. University of Chicago Press.

Matthew Smith. (2020). What do Americans think socialism looks like?

Mazzocchi, F. (2021). Being interconnected at the time of covid-19 pandemic: A call to regain the sense of community. *Journal of Futures Studies*, *25*(3), 39–48.

McElreath, R., Boyd, R., & Richerson, P. J. (2003). Shared norms and the evolution of ethnic markers. *Current Anthropology*, *44*(1), 122–130.

McGlynn, J., Baryshevtsev, M., & Dayton, Z. A. (2020). Misinformation more likely to use non-specific authority references: Twitter analysis of two covid-19 myths. *Harvard Kennedy School Misinformation Review*, *1*(3).

McGreal, C., Beckett, L., Laughland, O., & Ajasa, A. (2021). Derek chauvin found guilty of murder of george floyd. *The Guardian*.

McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., & Halgamuge, M. N. (2023). From google gemini to openai q*(q-star): A survey of reshaping the generative artificial intelligence (ai) research landscape. *arXiv preprint arXiv:2312.10868*.

Mehta, S. (2023). *Twitter's throne is for the taking as bluesky shows promise*. SAGE Publications: SAGE Business Cases Originals.

Mondal, M., Silva, L. A., Correa, D., & Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. *New Review of Hypermedia and Multimedia*, *24*(2), 110–130.

Morrow, G., Swire-Thompson, B., Polny, J. M., Kopec, M., & Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, *73*(10), 1365–1386.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A bert-based transfer learning approach for hate speech detection in online social media. *International Conference on Complex Networks and Their Applications*, 928–940.

Muhammed T, S., & Mathew, S. K. (2022). The disaster of misinformation: A review of research in social media. *International journal of data science and analytics*, *13*(4), 271–285.

Navarro, S. A., & Hernandez, S. L. (2022). *The color of covid-19: The racial inequality of marginalized communities*. Routledge.

Nesi, P., Pantaleo, G., Paoli, I., & Zaza, I. (2018). Assessing the retweet proneness of tweets: Predictive models for retweeting. *Multimedia Tools and Applications*, *77*(20), 26371–26396.

Nettle, D., & Dunbar, R. (1997). Social markers and the evolution of reciprocal exchange. *Current Anthropology*, *38*, 93–99.

Niven, D. (2021). Who says shut up and dribble? race and the response to athletes' political activism. *Journal of African American Studies*, *25*(2), 298–311.

Nsoesie, E. O., Cesare, N., Müller, M., & Ozonoff, A. (2020). Covid-19 misinformation spread in eight countries: Exponential growth modeling study. *Journal of medical Internet research*, *22*(12), e24425.

Orbe, M. (2015). #Alllivesmatter as post-racial rhetorical strategy. *Journal of Contemporary Rhetoric*, *5*.

ORTUTAY, B. (2023). The year of social media soul-searching: Twitter dies, x and threads are born and ai gets personal. *AP Online*, NA–NA.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, *115*(3), 999–1015.

Pan, Z., Lu, Y., Wang, B., & Chau, P. Y. (2017). Who do you think you are? common and differential effects of social self-identity on social media usage. *Journal of Management Information Systems*, *34*(1), 71–101.

Papadogiannakis, E., Papadopoulos, P., P. Markatos, E., & Kourtellis, N. (2023). Who funds misinformation? a systematic analysis of the ad-related profit routines of fake news sites. *Proceedings of the ACM Web Conference 2023*, 2765–2776.

Paul, J. (2019). 'not black and white, but black and red': Anti-identity identity politics and# alllivesmatter. *Ethnicities*, *19*(1), 3–19.

Pew Research Center. (2021). Beyond red vs. blue: The political typology.

Phillips, R. L. (2019). # Marchforourlives: Mobilization of a gun violence prevention movement on twitter.

Pierri, F., Perry, B. L., DeVerna, M. R., Yang, K.-C., Flammini, A., Menczer, F., & Bryden, J. (2022). Online misinformation is linked to early covid-19 vaccination hesitancy and refusal. *Scientific reports*, *12*(1), 5966.

Pratto, F., Çidam, A., Stewart, A. L., Zeineddine, F. B., Aranda, M., Aiello, A., Chryssochoou, X., Cichocka, A., Cohrs, J. C., Durrheim, K., et al. (2013). Social dominance in context and in individuals: Contextual moderation of robust effects of social dominance orientation in 15 languages and 20 countries. *Social Psychological and Personality Science*, *4*(5), 587–599.

Pyzik, O., Hertig, J., Kanso, H., Chamba, A., & Khan, S. (2021). The relationship between fake news and fake medicines: How misinformation has fuelled the sale of covid-19 substandard and falsified medical products. *Journal of EAHIL*, *17*(4), 17.

Radwan, N. (2022). The internet's role in undermining the credibility of the healthcare industry. *International Journal of Computations, Information and Manufacturing (IJCIM)*, *2*(1).

Reid, A., Ronda-Perez, E., & Schenker, M. B. (2021). Migrant workers, essential work, and covid-19. *American Journal of Industrial Medicine*, *64*(2), 73–77.

Reihani, H., Ghassemi, M., Mazer-Amirshahi, M., Aljohani, B., & Pourmand, A. (2021). Non-evidenced based treatment: An unintended cause of morbidity and mortality related to covid-19. *The American journal of emergency medicine*, *39*, 221.

Ricard, J., & Medeiros, J. (2020). Using misinformation as a political weapon: Covid-19 and bolsonaro in brazil. *Harvard Kennedy School Misinformation Review*, *1*(3).

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on twitter: A survey. *Information processing & management*, *52*(5), 949–975.

Rizeq, J., Flora, D. B., & Toplak, M. E. (2021). An examination of the underlying dimensional structure of three domains of contaminated mindware: Paranormal beliefs, conspiracy beliefs, and anti-science attitudes. *Thinking & Reasoning*, *27*(2), 187–211.

Röchert, D., Shahi, G. K., Neubaum, G., Ross, B., & Stieglitz, S. (2021). The networked context of covid-19 misinformation: Informational homogeneity on youtube at the beginning of the pandemic. *Online Social Networks and Media*, *26*, 100164.

Rovetta, A., & Bhagavathula, A. S. (2020). Covid-19-related web search behaviors and infodemic attitudes in italy: Infodemiological study. *JMIR public health and surveillance*, *6*(2), e19374.

Safarnejad, L., Xu, Q., Ge, Y., Krishnan, S., Bagarvathi, A., & Chen, S. (2020). Contrasting misinformation and real-information dissemination network structures on social media during a health emergency. *American journal of public health*, *110*(S3), S340–S347.

Shi, Y., Wang, G., Cai, X.-p., Deng, J.-w., Zheng, L., Zhu, H.-h., Zheng, M., Yang, B., & Chen, Z. (2020). An overview of covid-19. *Journal of Zhejiang University. Science. B*, *21*(5), 343.

Smaldino, P. E. (2019). Social identity and cooperation in cultural evolution. *Behavioural Processes*, *161*, 108–116.

Smaldino, P. E., Flamson, T. J., & McElreath, R. (2018). The evolution of covert signaling. *Scientific reports*, *8*(1), 1–10.

Smaldino, P. E., & Turner, M. A. (2021). Covert signaling is an adaptive communication strategy in diverse populations. *Psychological Review.*

Sosis, R., & Alcorta, C. (2003). Signaling, solidarity, and the sacred: The evolution of religious behavior. *Evolutionary Anthropology*, *12*(6), 264–274.

Strassle, P. D., Stewart, A. L., Quintero, S. M., Bonilla, J., Alhomsi, A., Santana-Ufret, V., Maldonado, A. I., Forde, A. T., & Nápoles, A. M. (2022). Covid-19–related discrimination among racial/ethnic minorities and other marginalized communities in the united states. *American Journal of Public Health*, *112*(3), 453–466.

Tabachnick, B., & Fidell, L. (2007). Using multivariate statistics., 5th edn.(pearson: Boston, ma.)

Tawa, J., Ma, R., & Katsumoto, S. (2016). “all lives matter”: The cost of colorblind racial attitudes in diverse social networks. *Race and Social Problems*, *8*(2), 196–208.

Tessler, H., Choi, M., & Kao, G. (2020). The anxiety of being asian american: Hate crimes and negative biases during the covid-19 pandemic. *American Journal of Criminal Justice*, *45*, 636–646.

Thomas, D., & Horowitz, J. M. (n.d.). Support for black lives matter has decreased since june but remains strong among black americans.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From louvain to leiden: Guaranteeing well-connected communities. *Scientific reports*, *9*(1), 1–12.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, *22*(3), 213–224.

Van der Linden, S. (2015). The conspiracy-effect: Exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. *Personality and Individual Differences*, *87*, 171–173.

van der Does, T., Galesic, M., Dunivin, Z. O., & Smaldino, P. E. (2022). Strategic identity signaling in heterogeneous networks. *Proceedings of the National Academy of Sciences*, *119*(10), e2117898119. https://doi.org/10.1073/pnas.2117898119

van Mulukom, V. (2020). The role of trust and information in adherence to protective behaviors & conspiracy thinking during the covid-19 pandemic and infodemic_preprint. *PsyArXiv*.

van Mulukom, V., Pummerer, L. J., Alper, S., Bai, H., Čavojová, V., Farias, J., Kay, C. S., Lazarevic, L. B., Lobato, E. J., Marinthe, G., et al. (2022). Antecedents and consequences of covid-19 conspiracy beliefs: A systematic review. *Social Science & Medicine*, 114912.

van Prooijen, J.-W., & Douglas, K. M. (2018). Belief in conspiracy theories: Basic principles of an emerging research domain. *European journal of social psychology*, *48*(7), 897–908.

Vegetti, F., & Littvay, L. (2021). Belief in conspiracy theories and attitudes toward political violence. *Italian Political Science Review/Rivista Italiana di Scienza Politica*, 1–15.

Voggeser, B. J., Singh, R. K., & Göritz, A. S. (2018). Self-control in online discussions: Disinhibited online behavior as a failure to recognize social cues. *Frontiers in psychology*, *8*, 2372.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wangersky, P. J. (1978). Lotka-volterra population models. *Annual Review of Ecology and Systematics*, *9*(1), 189–218.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. *Proceedings of the first workshop on NLP and computational social science*, 138–142.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. http://www.aclweb.org/anthology/N16-2013

West, K., Greenland, K., & van Laar, C. (2021). Implicit racism, colour blindness, and narrow definitions of discrimination: Why some white people prefer 'all lives matter'to 'black lives matter'. *British Journal of Social Psychology*, *60*(4), 1136–1153.

Whittaker, E., & Kowalski, R. M. (2015). Cyberbullying via social media. *Journal of School Violence*, *14*(1), 11–29.

Wood, M. J., Douglas, K. M., & Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science*, *3*(6), 767–773.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136.

Yamey, G., & Gonsalves, G. (2020). Donald trump: A political determinant of covid-19.

Yılmaz, S. R. (2017). The role of social media activism in new social movements: Opportunities and limitations. *International Journal of Social Inquiry*, *10*(1), 141–164.

Young, D. G., Bagozzi, B. E., Goldring, A., Poulsen, S., & Drouin, E. (2019). Psychology, political ideology, and humor appreciation: Why is satire so liberal? *Psychology of Popular Media Culture*, *8*(2), 134.

Zimmerman, A. G., & Ybarra, G. J. (2016). Online aggression: The influences of anonymity and social modeling. *Psychology of Popular Media Culture*, *5*(2), 181.

Appendix A

Appendix

A.1 Say Their Names

SONYA MASSEY AKIRA ROSS JORDAN NEELY DARRYL TYREE WILLIAMS TYRE NICHOLS KEENAN ANDERSON SINZAE REED KESHAWN THOMAS DANTE KITTRELL JAYLAND WALKER CHRISTOPHER KELLEY RUTH WHITFIELD PEARL YOUNG KATHERINE MASSEY HEYWARD PATTERSON CELESTINE CHANEY GERALDINE TALLEY AARON SALTER JR. ANDRE MACKNIEL MARGUS MORRISON ROBERTA DRURY PATRICK LYOYA DONNELL ROCHESTER AMIR LOCKE ISAIAH TYREE WILLIAMS JASON WALKER JAMES WILLIAMS MICHAEL WAYNE JACKSON ARNELL "AJ" STEWART FANTA BILITY ALVIN MOTLEY JR. TA'NEASHA CHAPPELL RYAN LEROUX WINSTON SMITH LATOYA DENISE JAMES ANDREW BROWN JR. MA'KHIA BRYANT MATTHEW "ZADOK" WILLIAMS DAUNTE WRIGHT JAMES LIONEL JOHNSON DOMINIQUE WILLIAMS DONOVON LYNCH MARVIN SCOTT III JENOAH DONALD PATRICK WARREN XZAVIER HILL ROBERT HOWARD VINCENT BELMONTE MONICA GOODS BENNIE EDWARDS CASEY GOODSON JR. AIDEN ELLISON QUAWAN CHARLES KEVIN PETERSON JR. WALTER WALLACE JR. JONATHAN PRICE KURT REINHOLD DIJON KIZZEE DAMIAN DANIELS ANTHONY MCCLAIN JULIAN LEWIS MAURICE ABISDID-WAGNER BRAYLA STONE RAYSHARD BROOKS PRISCILLA SLATER ROBERT FORBES KAMAL FLOWERS JAMEL FLOYD DAVID MCATEE JAMES SCURLOCK CALVIN HORTON JR. TONY MCDADE DION JOHNSON GEORGE FLOYD MAURICE GORDON CORNELIUS FREDERICKS STEVEN TAYLOR DANIEL PRUDE BREONNA TAYLOR BARRY GEDEUS MANUEL ELLIS REGINALD "REGGIE" PAYNE AHMAUD ARBERY LIONEL MORRIS JAQUYN O'NEILL LIGHT WILLIAM GREEN

DARIUS TARVER   MICIAH LEE   JOHN NEVILLE   CAMERON LAMB   MICHAEL DEAN   ATATIANA JEFFERSON   BYRON WILLIAMS   ELIJAH MCCLAIN   JALEEL MEDLOCK   TITI "TETE" GULLEY   DOMINIQUE CLAYTON   PAMELA TURNER   RONALD GREENE   STERLING HIGGINS   BRADLEY BLACKSHIRE   JASSMINE MCBRIDE   ALEAH JENKINS   EMANTIC BRADFORD JR.   JEMEL ROBERSON   CHARLES ROUNDTREE JR.   BOTHAM JEAN   HARITH AUGUSTUS   JASON WASHINGTON   ANTWON ROSE JR.   ROBERT WHITE   EARL MCNEIL   MARCUS-DAVID PETERS   DORIAN HARRIS   DANNY RAY THOMAS   STEPHON CLARK   RONELL FOSTER   DAMON GRIMES   JAMES LACY   CHARLEENA LYLES   MIKEL MCINTYRE   JORDAN EDWARDS   TIMOTHY CAUGHMAN   ALTERIA WOODS   DESMOND PHILLIPS   DEBORAH DANNER   ALFRED OLANGO   TERENCE CRUTCHER   CHRISTIAN TAYLOR   JAMARION ROBINSON   DONNELL THOMPSON JR.   JOSEPH MANN   PHILANDO CASTILE   ALTON STERLING   JAY ANDERSON JR.   CHE TAYLOR   DAVID JOSEPH   ANTRONIE SCOTT   BETTIE JONES   QUINTONIO LEGRIER   COREY JONES   SAMUEL DUBOSE   DARRIUS STEWART   SANDRA BLAND   SUSIE JACKSON   DANIEL SIMMONS   ETHEL LANCE   MYRA THOMPSON   CYNTHIA HURD   DEPAYNE MIDDLETON-DOCTOR   SHARONDA COLEMAN-SINGLETON   CLEMENTA PINCKNEY   TYWANZA SANDERS   KALIEF BROWDER   FREDDIE GRAY   NORMAN COOPER   WALTER SCOTT   ERIC HARRIS   MEAGAN HOCKADAY   NATASHA MCKENNA   RUMAIN BRISBON   TAMIR RICE   AKAI GURLEY   TANISHA ANDERSON   LAQUAN MCDONALD   CAMERON TILLMAN   DARRIEN HUNT   KAJIEME POWELL   MICHELLE CUSSEAUX   DANTE PARKER   EZELL FORD   MICHAEL BROWN   AMIR BROOKS   JOHN CRAWFORD III   ERIC GARNER   JERRY DWIGHT BROWN   VICTOR WHITE III   MARQUISE JONES   YVETTE SMITH   RENISHA MCBRIDE   JONATHAN FERRELL   DEION FLUDD   GABRIEL WINZER   WAYNE A. JONES   KIMANI GRAY   KAYLA MOORE   COREY STINGLEY   DARNESHA HARRIS   JORDAN DAVIS   MOHAMED BAH   SGT. JAMES BROWN   DARIUS SIMMONS   REKIA BOYD   TRAYVON MARTIN   WILLIE RAY BANKS   KENNETH CHAMBERLAIN SR.   CLETIS WILLIAMS   ROBERT RICKS   EUGENE ELLISON   DANROY "DJ" HENRY JR.   AIYANA STANLEY-JONES   LAWRENCE ALLEN   OSCAR GRANT   JULIAN ALEXANDER   MARVIN PARKER   DEAUNTA FARROW   SEAN BELL   KATHRYN JOHNSTON   TIMOTHY STANSBURY JR.   ALBERTA SPRUILL   ANTHONY DWAIN LEE   RICKY BYRDSONG   AMADOU DIALLO   JAMES BYRD JR.   NICHOLAS HEYWARD JR.   MARY MITCHELL   SHARON WALKER   ELEANOR BUMPURS   EDWARD GARDNER   ELTON HAYES   FRED HAMPTON   MARTIN LUTHER KING JR.   ALBERTA ODELL JONES   JIMMIE LEE JACKSON   MALCOLM X   JAMES EARL CHANEY   LOUIS ALLEN   MEDGAR EVERS   HERBERT LEE   JOHN EARL REESE   EMMETT TILL   WILLIAM

MCDUFFIE DELLA MCDUFFIE MALCOLM WRIGHT GEORGE STINNEY JR.
DR. ANDREW C. JACKSON WILL BROWN LEVI HARRINGTON

A.2 Hashtags as Signals of Political Identity

Conservative Response	**Liberal Response**
Homosexuality should be discouraged by society.	Homosexuality should be accepted by society.
Stricter environmental laws and regulations cost too many jobs and hurt the economy.	Stricter environmental laws and regulations are worth the cost.
Most corporations make a fair and reasonable amount of profit.	Business corporations make too much profit.
The best way to ensure peace is through military strength.	Good diplomacy is the best way to ensure peace.
Immigrants today are a burden on our country because they take our jobs, housing, and health care.	Immigrants today strengthen our country because of their hard work and talents.
Black people who can't get ahead in this country are mostly responsible for their own condition.	Racial discrimination is the main reason why many Black people can't get ahead these days.
The government today can't afford to do much more to help the needy.	The government should do more to help needy Americans, even if it means going deeper into debt.
Poor people today have it easy because they can get government benefits without doing anything in return.	Poor people have hard lives because government benefits don't go far enough to help them live decently.
Government regulation of business usually does more harm than good.	Government regulation of business is necessary to protect the public.
Government is almost always wasteful and inefficient.	Government often does a better job than people give it credit for.

Table A.1: “Conservative" and “Liberal" responses used to compute Political Orientation scores.

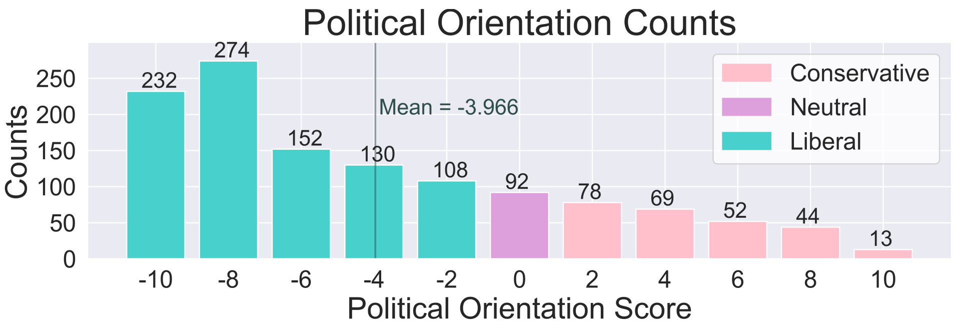


Figure A.1: Distribution of political orientation scores with $\mu = -3.966$.

	Question	Responses
COR 0	Do you consider yourself to be religious?	Yes No
COR 1	How often do you take part in religious services?	More than once a week Once a week One to three times a month A few times a year Less often Never
COR 2	How important is it to take part in religious services?	Very much so Quite a bit Moderately Not very much Not at all
COR 3	How important is it for you to be connected to a religious community?	

Table A.2: Centrality of religiosity questions and possible responses (Huber & Huber, 2012).

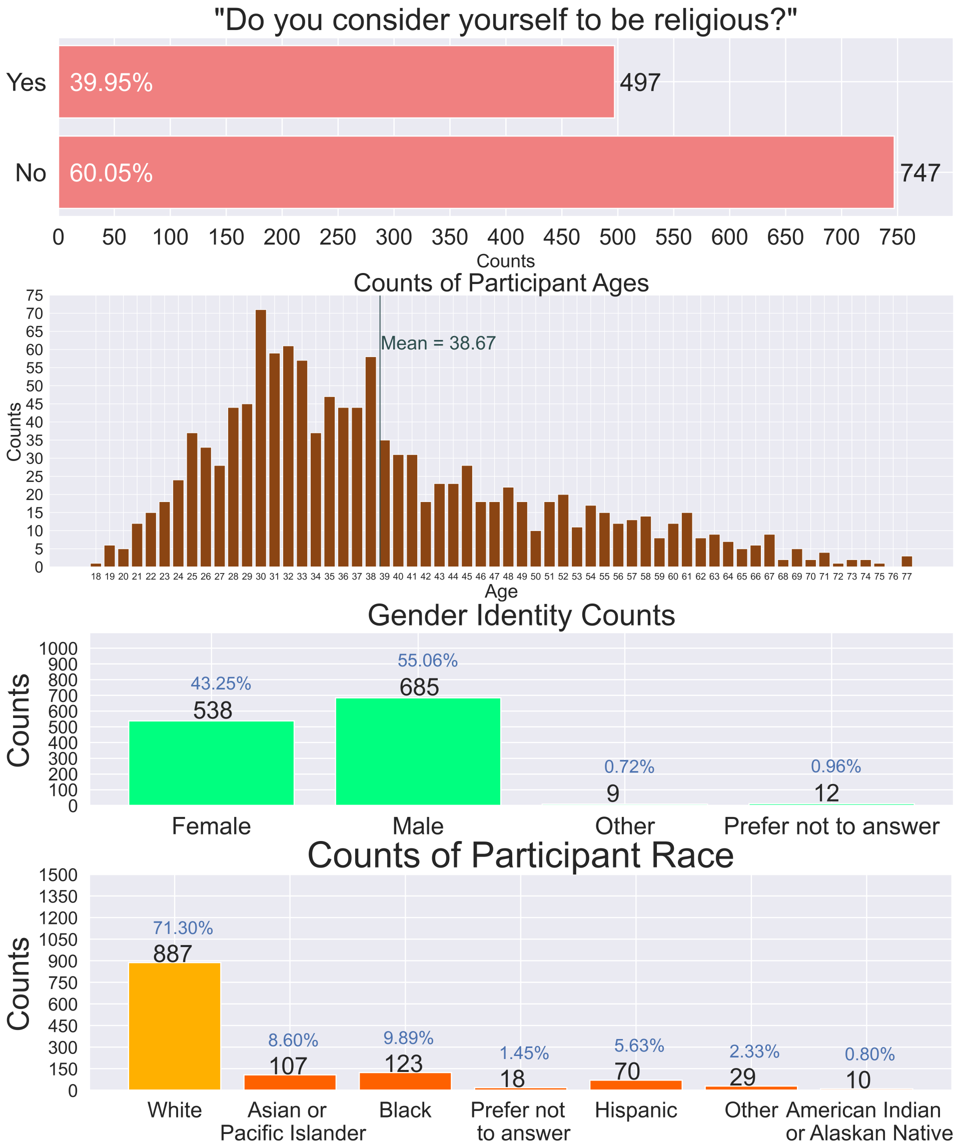


Figure A.2: Participant responses to Centrality of Religiosity question 0. Participant distribution of ages with mean 39. Participant distribution of gender identities. Participant distribution of race.

	Check Question	Responses
1	Please select "strongly agree".	Strongly disagree Somewhat disagree Neither agree nor disagree Somewhat agree Agree Strongly agree
2	Which of the following is NOT an animal?	Chair Cow Cat

Table A.3: Check questions included in the survey to gauge user attentiveness.

Model	Response	Multivariate Linear Regression	Random Forests
0	#AllLivesMatter with hashtag for “racist”	0.1446	0.1771
1	#AllLivesMatter without hashtag for “racist”	0.08475	0.0985
2	#BlackLivesMatter with hashtag for “racist”	0.2194	0.2220
3	#BlackLivesMatter without hashtag for “racist”	0.161	0.1822
4	#AllLivesMatter with hashtag for “offensive”	0.232	0.2583
5	#AllLivesMatter without hashtag for “offensive”	0.09915	0.1117
6	#BlackLivesMatter with hashtag for “offensive”	0.2599	0.2534
7	#BlackLivesMatter without hashtag for “offensive”	0.249	0.2615

Table A.4: R^2 values for each of the 8 full models using Multivariate Linear Regression and Random Forests.

Model	Response	f-statistic	p-value
0	#AllLivesMatter with hashtag for “racist”	84.7433	5.6035×10^{-19}
1	#AllLivesMatter without hashtag for “racist”	28.9448	1.0659×10^{-7}
2	#BlackLivesMatter with hashtag for “racist”	83.5795	9.3898×10^{-19}
3	#BlackLivesMatter without hashtag for “racist”	22.8853	2.1629×10^{-6}
4	#AllLivesMatter with hashtag for “offensive”	130.8381	1.5205×10^{-27}
5	#AllLivesMatter without hashtag for “offensive”	52.3483	1.4134×10^{-12}
6	#BlackLivesMatter with hashtag for “offensive”	128.0722	4.7837×10^{-27}
7	#BlackLivesMatter without hashtag for “offensive”	85.4679	3.9786×10^{-19}

Table A.5: Results from partial f-test analysis on each of the 8 models that yield the largest f-statistics. The nested models for all of these results include all variables *except* political orientation score.

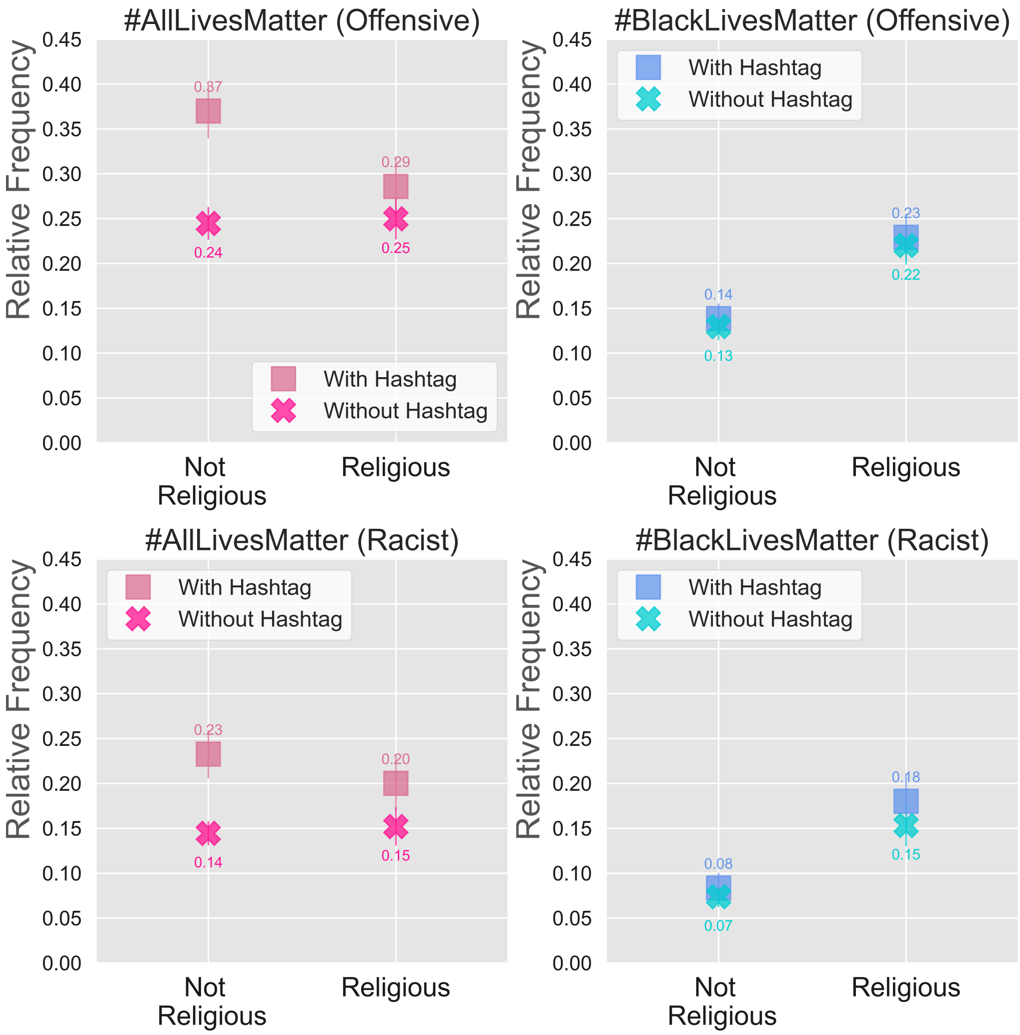


Figure A.3: Evaluations of #AllLivesMatter and #BlackLivesMatter tweets by self- identification of religiosity (Centrality of Religion Question 0) with 95% confidence interval, where relative frequency is calculated by dividing racist or offensive counts by total counts. Independent t-tests revealed that there were statistically significant differences between evaluations of #AllLivesMatter tweets with hashtags present versus without, with the strongest effect present in evaluations of tweets as offensive (corresponding p-values of 1.514×10^{-11} and 0.070 for non religious and religious participant evaluations, respectively) followed by evaluations of tweets as racist (corresponding p-values of 2.605×10^{-8} and 0.010 for non religious and religious participant evaluations, respectively). Differences of evaluations of #BlackLivesMatter tweets with hashtags present versus without had much weaker effects, with offensive ratings (corresponding p-values of 0.442 and 0.598 for non religious and religious participant evaluations, respectively) having a slightly weaker effect than racist ratings (corresponding p-values of 0.366 and 0.164 for non religious and religious participant evaluations, respectively).

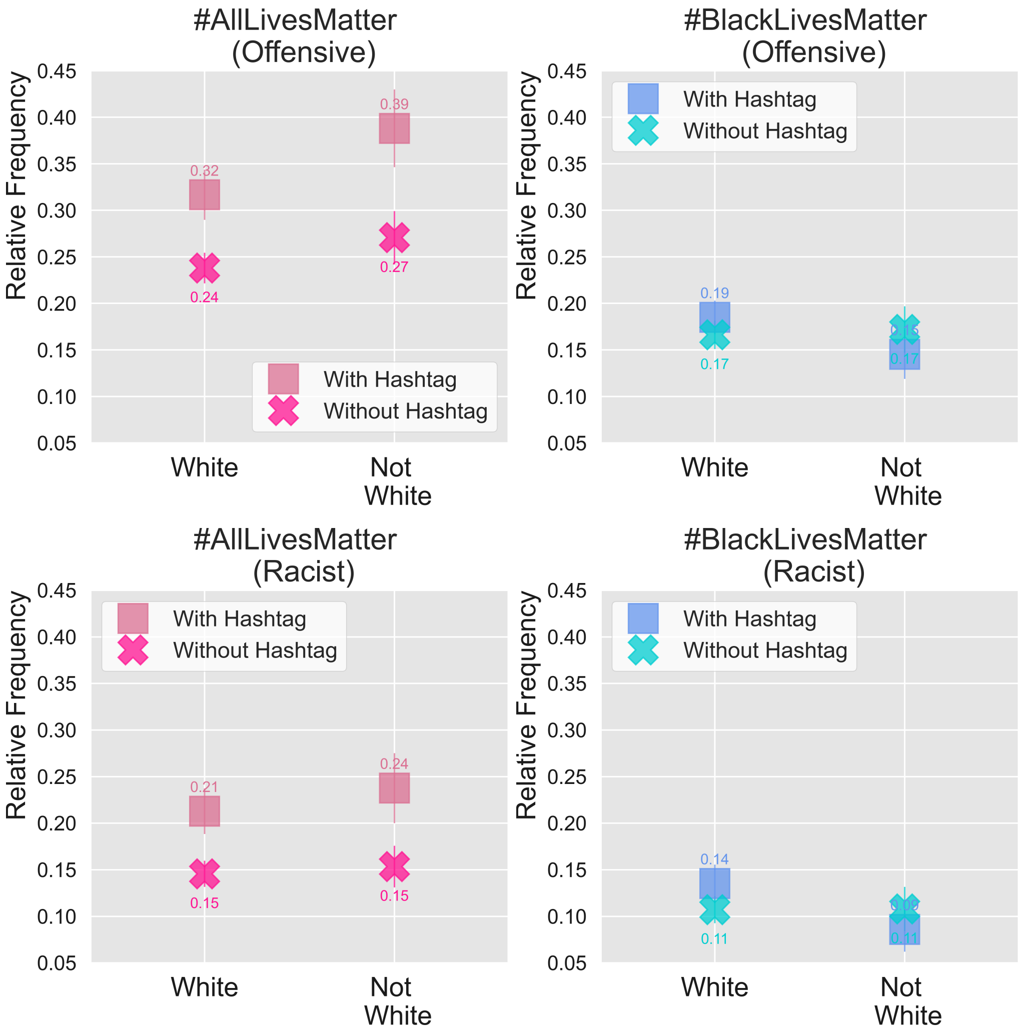


Figure A.4: Evaluations of #AllLivesMatter and #BlackLivesMatter tweets by white v. non-white participants with 95% confidence interval, where relative frequency is calculated by dividing racist or offensive counts by total counts. Independent t-tests revealed that there were statistically significant differences between evaluations of #AllLivesMatter tweets with hashtags present versus without, with the strongest effect present in evaluations of tweets as offensive (corresponding p-values of 8.976×10^{-7} and 8.967×10^{-6} for white and not white participant evaluations, respectively) followed by evaluations of tweets as racist (corresponding p-values of 2.456×10^{-6} and 0.0002 for white and not white participant evaluations, respectively). Differences of evaluations of #BlackLivesMatter tweets with hashtags present versus without had much weaker effects, with offensive ratings (corresponding p-values of 0.116 and 0.151 for non religious and religious participant evaluations, respectively) having a slightly stronger effect than racist ratings (corresponding p-values of 0.027 and 0.198 for non religious and religious participant evaluations, respectively).

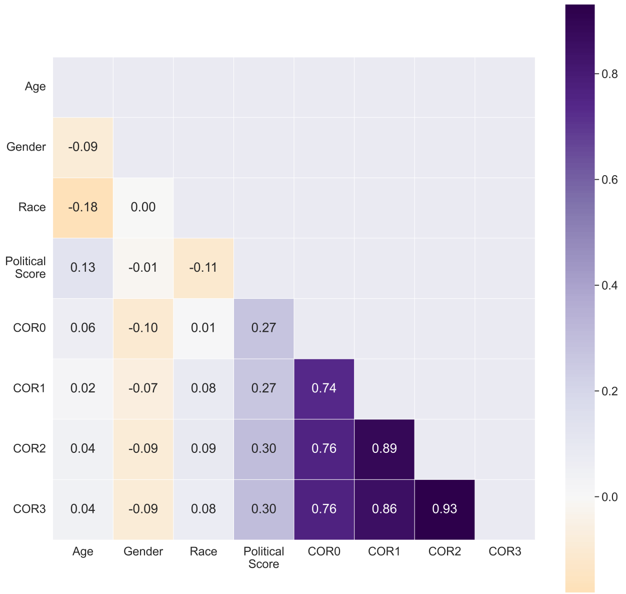


Figure A.5: Correlation between select demographics (gender, race, political score, and religiosity question results).

Dataset	Racist						Offensive					
	ICC(2,k)			ICC(3,k)			ICC(2,k)			ICC(3,k)		
	ICC	F-statistic	p-value	ICC	F-statistic	p-value	ICC	F-statistic	p-value	ICC	F-statistic	p-value
1	0.928	16.533	5.63E-77	0.94	16.533	5.63E-77	0.951	22.889	1.57E-109	0.956	22.889	1.57E-109
2	0.962	29.889	7.57E-146	0.967	29.889	7.57E-146	0.98	56.021	1.62E-266	0.982	56.021	1.62E-266
3	0.9	12.275	6.18E-55	0.919	12.275	6.18E-55	0.915	12.685	4.13E-57	0.921	12.685	4.13E-57
4	0.958	31.145	3.17E-151	0.968	31.145	3.17E-151	0.956	25.095	2.76E-121	0.96	25.095	2.76E-121
5	0.967	37.11	1.39E-179	0.973	37.11	1.39E-179	0.967	32.954	7.95E-160	0.97	32.954	7.95E-160
6	0.96	29.656	6.01E-144	0.966	29.656	6.01E-144	0.956	25.251	4.57E-122	0.96	25.251	4.57E-122
7	0.873	8.998	1.86E-37	0.889	8.998	1.86E-37	0.924	15.669	1.76E-72	0.936	15.669	1.76E-72
8	0.971	41.51	9.10E-200	0.976	41.51	9.10E-200	0.974	41.958	8.25E-202	0.976	41.958	8.25E-202
9	0.966	33.25	4.37E-161	0.97	33.25	4.37E-161	0.962	30.036	1.53E-145	0.967	30.036	1.53E-145
10	0.97	38.556	3.65E-187	0.974	38.556	3.65E-187	0.969	36.396	5.88E-177	0.973	36.396	5.88E-177

Table A.6: Relevant values for intraclass correlation tests.