Online Political Communication

Persuasion in a Hostile Place

by

Kevin Munger

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Politics

New York University

September 2018

Jonathan Nagler

For my parents

Acknowledgments

This work would not have been possible without the support and intellectual contribution of a whole host of individuals, to each of whom I am deeply indebted. I can't hope to name them all, but I want to specifically thank:

my dissertation committee, Joshua Tucker, Patrick Egan, and especially my dissertation advisor Jonathan Nagler, who have been essential in developing and encouraging my ambition to conduct the research presented in this dissertation, and my dissertation readers Chris Dawes and James Hamilton, for whose expertise I am grateful;

all of the participants, programmers, and PIs of the NYU Social Media and Political Participation (SMaPP), who provided a stimulating intellectual environment where the significance of social media for politics was never in doubt—and in particular, Pablo Barberà, whose work with the SMaPP lab and since has been so influential;

other members of the NYU Politics community, especially professors Neal Beck, Cyrus Samii, Jennifer Larson, Melissa Schwartzberg, Eric Dickson, and Rebecca Morton, and my former and present graduate colleagues, especially Leevio di Lonardo, Hodgdon Bisbee, Dim Drewery, Mateo Vasquez, Alex Siegel, Gabor Simonovits, Steve Rashin, Pedro Rodriguez, Abraham Aldama, Denis Stukal, Jason Guo, Saad Gulzar, Hannah Simpson, Carlo Horz,

Renard Sexton, Mai Nguyen, Antonella Bandiera, Maria Carreri, and Sean Kates, all of whom provided valuable feedback and support;

the organizers and participants of the conferences and seminar series where I presented earlier versions of this work, including the NYU Center for Experimental Social Science Conference, the Toronto Political Behaviour Workshop, the MIT Online Harassment Workshop, the Crowdsourcing and Online Behavioral Experiments Workshop, the Harvard Experimental Political Science Graduate Conference, the Columbia Computational Social Science Working Group, the Yale Human Nature Lab, the Yale ISPS Experiments Workshop, and seminars at Kings College London, SUNY Stony Brook, University of Southern California, and Universidad de Rosario;

and my father, who has been a truly invaluable advisor throughout my life, and mother, without whom I don't think I'd be able to tie my shoes, much less write this dissertation.

Preface

"Communications tools don't get socially interesting until they get technologically boring...It's when a technology becomes so normal, then ubiquitous, and finally so pervasive as to be invisible, that the really profound changes happen"(Shirky, 2008).

The internet and social media are changing politics. For a variety of sociological reasons, academic Political Science had been slow to appreciate this fact, but the election of Donald Trump revealed the hollowing out of traditional media by the internet and the central role that attention plays in contemporary politics. Much of the research on this topic has focused on macro-level elite phenomena: social media platforms, online election campaigns, digitally-organized protests and the concomitant repression, censorship and misinformation campaigns.

This dissertation aims to deepen our understanding of the way that individuals engage with political content online, and how those behaviors might be changed. Relative to older media technologies, the internet is unique in the heterogeneity of uses it affords: there are many orders of magnitude more political news and opinions readily accessible online than in print media, television or radio. This is only possible because the reduction in the

costs of sharing information has enabled non-elites to broadcast their views at scale. The networked, algorithmic structure of social media means that an individual's decision to post has implications for the consumption menu available to everyone else.

Online political communication thus poses a challenge for democratic politics. Our democratic institutions—formal government structures, informal traditions, the mass media, academia—were all born well before the internet (as were all of the individual humans in charge of these institutions). These institutions will need to adapt, but seeing as they are democratic, this adaptation should (however imperfectly) be in response to the will of the people.

The more fundamental problem, then, is that people are not adapted to the logic of the internet. Norms are essential for structuring all human behavior, but norm enforcement online is made difficult by the absence of the evolved social and psychological stimuli that undergird real-world interactions. This has caused a proliferation of toxic online communication. Chapters one and two in this dissertation study the mechanisms of online norm enforcement.

The economics of online media represent a dramatic break from previous economic structures; the radically decreased costs have led to a proliferation of media companies, driving profits down in a race to the bottom for reader attention. This media system—which I call "clickbait"—is ultimately driven by consumer preferences between competing news stories. The Introduction outlines the theoretical foundations of clickbait media, while Chapter three presents the results of a series of experiments testing the implications of these preferences for trust in media and partisan polarization.

Kevin Munger: *Online Political Communication: Persuasion in a Hostile Place*

Advisor: Jonathan Nagler

Abstract

This dissertation analyzes the causes and moderators of three online political behaviors: racist harassment, partisan incivilty, and the consumption of partisan "clickbait" news. The technological affordances of the internet and social media have made it difficult to enforce deliberative speech norms, but have instead enabled the participation-depressing harassment of minorities. The first two chapters present the results of two online field experiments I conducted to study the promotion of healthier speech norms on Twitter. By varying the identity and the social status of the accounts I used to sanction subjects, I find that messages from high-status in-group accounts are more effective at changing behavior. By varying the content of the messages, I find that moral appeals are effective at changing behavior, while a non-moral message is not. The third chapter develops a theory of the economics of online media and provides evidence for the behavioral micro-foundation of "clickbait" media.

Contents

List of Figures

List of Tables

List of Appendices

Introduction: Clickbait Media

0.1 Introduction

The 2016 US Presidential election revealed that the mainstream news media are in a dire place. Trust in the institution has been declining steadily ever since the 1970s (Ladd, 2011), especially among conservatives. This latter trend has driven the development of a parallel conservative news media defined by cults of personality and outrage (Berry and Sobieraj, 2013).

The discussion of "Fake News" as a significant factor in the 2016 election is a striking manifestation of the lack of confidence in the media. Although current estimates place the number of people who recalled seeing the average "Fake News" story at 1.2% (Allcott and Gentzkow, 2017), that the concept resonated so well across both sides of the political spectrum indicates that people fear that the media can no longer perform its role as "gatekeeper," deciding which stories to disseminate to the public. Estimates of exposure to "Fake News" are considerably higher than of recall, and hugely moderated by age: Guess, Nyhan and Reifler (2017) find that"only 7% of people age 59 or younger consumed one or more pro-Trump Fake News articles compared with 19% among those age 60 or older."

The existence of "Fake News" is directly implied by the technological, regulatory and cultural structure of the online media industry; indeed, "Fake News" is less of an innovation than the culmination of existing trends in online media. In this chapter, I explore how the structure of the online media industry has evolved over the lifespan of the internet: entry into the market is cheaper; the costs of the inputs of online news production (information, skilled labor, distribution) are lower; consumers of political news are increasingly partisan and have widely varying levels of technical sophistication. These factors combine to produce what I call "credibility cascades", the mechanism which drives the online news industry: stories acquire credibility as they are shared along social networks, becoming more desirable at the same time as they increase their potential audience.

Underlying this macro-level phenomenon is the essential fact that each news story stands alone, competing for attention with all of the other news stories on the internet. In contrast to subscription-based newspapers (or magazines, or even websites), or to temporally linear formats like radio or television, media outlets creating content to be shared on Facebook are unable to bundle their news stories. The resulting pressure to attract attention affects both the format and subject matter of stories. During the heyday of nonpartisan media in the 20th century, journalists explicitly embraced their roles as curators and verifiers of information, roles which have today been subsumed by the structural logic of credibility cascades.

I call this media environment "Clickbait Media," which in its current iteration is better termed "Facebook Media." The importance of Facebook in particular to the contemporary media environment cannot be overstated; specific content and design-related decisions the company undertakes have the capacity to annihilate media companies around the globe.

In October 2017, for example, Facebook rolled out an experimental redesign to its central Newsfeed feature that separated posts from individual users and posts by "pages" (like those run by media companies) into separate feeds. Local media companies in the targeted countries—including Slovakia and Cambodia—reported drops of up to 75% in their audience size, making their business models untenable (Cellan-Jones, 2017). The modern media environment has been subject to Facebook's evolutionary pressures for years now, as the case of Upworthy illustrates.

This chapter summarizes the history of the media industry in the US to demonstrate that, in many ways, the industry has normally been structured like Clickbait Media; the cultural and technological conditions of the postwar era actually make *that* period the anomaly. I then describe the economic incentive structures that characterize Clickbait Media, as well as the specific cultural and technological conditions that make its current iteration—Facebook Media—distinct from earlier iterations. Finally, I describe survey data that tests the microfoundations of Clickbait Media. I find that certain types of people have a higher preference for clickbait—in particular, older people, people who use Facebook more often and read more news stories online, and moderate Republicans.

The survey data also includes several experiments which aim to demonstrate the implications of Clickbait Media for normative topics of contemporary importance: affective polarization, trust in media, and factual knowledge gained from political news online.

- *Affective Polarization*: The extent to which partisans feel positive towards co-partisans and negative towards opposite-partisans (Iyengar, Sood and Lelkes, 2012). The most straightforward way to measure the phenomenon is to look at differences between in-

group and out-group feeling thermometer measures.

- *Trust in Media*: Once seen as a neutral source of factual information, the media has suffered an erosion of reputation (Ladd, 2011). This tendency has been especially pronounced on the right, as conservative politicians since George McGovern have accused the media of left-wing bias and promoted an alternative knowledge structure (Grossmann and Hopkins, 2016). I ask people directly about how much they trust "online media" and "traditional media."

- *Factual Knowledge Gained from Political News Online*: The 2016 Election saw a frenzy of concern around "Fake News" spread on social media (Silverman, 2016). A crucial part of the story is that people update their factual beliefs based on what they see on social media (Munger, Egan, Nagler, Ronen and Tucker, 2016). I test whether people who read clickbait stories are more likely to remember facts from those stories.

The current experiments find robustly null effects of being randomly assigned to read clickbait headlines on any of these normative topics. I discuss the limitations of the current experimental setup and potential extensions in the conclusion.

0.2 A Brief History of Media in the United States

Facebook Media represents more of a return to business-as-usual than an unprecedented development. The baseline for media normalcy in our cultural consciousness is deceptive. The mid-1960s were the golden age of the institutional news media: a combination of the high barriers to entry but extensive reach of broadcast television, strict regulations about content,

and low partisan polarization allowed the news media to become fully institutionalized and commit to norms of professionalism and objectivity (Ladd, 2011). The rise of outrage media and the internet mirror the rise of cheap newsprint and yellow journalism a century ago.[1]

Prior to Hearst and Pulitzer, newspapers were primarily financed directly by political parties (or even the US government). Objectivity was not generally seen as essential or even necessarily desirable. The proliferation of high-volume printing presses in the late 19th century changed the business model of publishers in large cities, as subscription models waned in favor of on-the-spot sales. As a result, headlines became designed to grab as much attention as possible. There were even some papers that featured quizzes on the front page that purported to tell about the reader's personality and predict their future. There truly is nothing new under the sun—not even Buzzfeed.

This structural shift made newspapers more entertainment-focused, and part of that entertainment meant taking political positions. The big change, though, was *away* from being officially associated with major parties. So the papers took up populist issues that appealed to their readers' sensibilities, sometimes at the expense of sound reasoning or sober policy advice—most famously, the incitement to war against the Spanish in 1898.

This media-economics state was one in a which a technology (newsprint) developed to the point where the marginal cost of production was driven to zero, and in which opting into reading any particular article (as hawked by a newsboy, at least in cities) could be done at a whim. I claim that this represents the mature stage of a given medium, in which competition and technological innovation reduce costs and give consumers maximal choice.

[1] Much of the following section is developed in more detail in Ladd's excellent *Why Americans Hate the Media, and Why it Matters*.

Radio and television followed the same pattern, from a near-monopoly due to high startup costs to the eventual democratization of production. These technologies were different from the Yellow Journalism equilibrium of the turn of the century, though, because of their fundamentally linear nature: unlike newspapers, consumption could not be begun or ended on demand, limiting the actual choice available to the consumer. Interestingly, we may be seeing the final maturation of these forms only because of the internet, which allows radio and television to be consumed on demand. As Mutz (2015) points out, the internet as a distribution platform reduces the mechanical distinctions between these mediums.

The period when broadcast television was the primary outlet for news represents *an anomaly* in the economics of media; the proliferation of cable news channels, talk radio stations, blogs and social media seem to be the norm given mature technologies and a moderate regulatory burden.

However, the current cultural and technological environment produces a novel media industry-political news dynamic: contemporary Clickbait Media is Facebook Media. The term "clickbait" is clearly native to the computer; Merriam-Webster, which added the word to its dictionary in 2015, defines it as "something (such as a headline) designed to make readers want to click on a hyperlink especially when the link leads to content of dubious value or interest," and claims that it was first used in 2010 (Editors, 2015).

Clickbait is the essential manifestation of Clickbait Media, but the specific form it takes has varied over time, even in the brief period that the term has existed. The example of Upworthy is illustrative. The "fastest-growing media site of all time," Upworthy implemented a new style of clickbait headline designed to entice consumption by strategically withholding

information (Sanders, 2017). Less than two years after it was founded, Upworthy had over 80 million unique visitors each month–more than either the New York Times or Washington Post. In November 2013, however, Facebook announced that it would penalize deceptive headlines in their ranking algorithm, and within a year, Upworthy's business collapsed. In November 2014, the site had only 20 million unique visitors (Sanders, 2017). These changes had massive implications for how people consumed political news in the early 2010s, as other sites quickly caught on to Upworthy's success.

The flood of news sites inspired by Upworthy was enabled and entailed by the novel economic incentive structure of Facebook Media. The rapid rise and fall of Upworthy points to the necessity of understanding that incentive structure to appreciate which elements of Facebook Media are central and which are ephemeral. The following section aims to adapt the approach in **?** to account for the way that new media technologies have altered the market for news. I leverage an old and well developed theoretical framework from economics that applies remarkably well to Facebook Media: contestable markets (Baumol et al., 1982).

0.3 The Economics of Facebook Media: Contestable Markets

Motivated by the example of the newly-deregulated airline industry, William Baumol and his coauthors described several economic conditions that would need to obtain to create a peculiar industry structure. It was cheap to rent an airplane and operate a very small-scale business. This operation would easily be able to out-compete existing firms, with their large organizational and legacy costs; airline flights are a relatively undifferentiated product, so price is the main way these firms compete. Baumol called this a *contestable market*, and

made several predictions about how firms might behave in such an industry.

In the short run, new firms would engage in "hit-and-run" competition, entering the industry and charging sub-market rates in order to attract business from the established firms. Once the market price has adjusted to account for the increase in supply, these new firms will exit the industry. In the long run, then, there will be very few firms in the industry, each of them earning close to zero profit. The threat of entry prevents any firm from charging a price higher than their marginal cost in order to earn a profit.

Baumol identified several aspects necessary to characterize a contestable market. In the real world in which entropy exists, none of these idealized conditions can actually occur, so it makes more sense to refer to the degrees of contestability in a market rather than contestability as a binary condition.

- Entry/Exit: There cannot be formal barriers to entry or exit from the market. Firms need to be able to rapidly set up shop without any kind of explicit licensing requirements, and they need to be able to easily exit the market–importantly, they cannot have long-term commitments like employee pensions.
- Sunk Costs: There cannot be any capital investments (in either physical or intellectual capital) that cannot be recouped.
- Symmetric Information/Technology: There cannot be any specialized technology or knowledge available to the incumbent firms but not the new entrants.

Baumol may have been wrong about the airline industry being a contestable market; at least, it failed to be characterized by rapid entry and exit of extremely low-cost firms (Martin, 2000). However, the three above conditions *do* describe contemporary Facebook Media. We

appear to be in the short term "hit-and-run" competition described by Baumol: the barriers to entry for new firms have been decreasing, the sunk costs associated with a media firm have been disappearing, and the primary technology of news creation—access to information—has been radically democratized. The current environment sees more (national) media firms directly competing with each other than ever before in history.

The crucial question is what will happen in the long run. The current situation does not appear sustainable, and indeed contestable markets theory predicts that it is not. The new entrants are playing a different game than legacy media outlets, producing online content at a fraction of the price. It may be the case that the steady state of this market is also what constable markets theory predicts about the long-run equilibrium: very few firms making very low profits, constrained by the threat of other costless entrants. However, real-world conditions necessarily deviate from idealized economic theory. My theory of Facebook Media describes the current state of online news media to explain how it deviates from an ideally contestable market and offers several alternative possible equilibria for the industry.

0.3.1 Entry/Exit

Until recently, it was very costly to create a national news organization. The logistical challenges of distributing a physical newspaper meant that from WWII to today, only one national newspaper (the USA Today, in 1982) entered the marketplace (Hindman, 2008). The broadcast television market was limited to three firms by a combination of regulation and massive capital requirements. The cable television market was less regulated and somewhat cheaper to enter; several new cable television firms were founded, but costs and regulations limited this number as well(Prior, 2007).

The spread of the internet (and the necessary hardware and software) allowed anyone with some degree of technical know-how to set up their own blog and create news content on their own. These blogs allowed many people to share their views outside the supervision of traditioal media companies, but initial enthusiasm about the democratization of information production was misplaced. Due to the link structure of the early web, attention was distributed according to what Hindman (2008) calls the "Googlearchy," and all but the very most successful blogs received next to zero attention.

The online media market only became mature as internet use became common among a much broader swath of the population, a trend which co-developed with the ubiquity of social media use. In 2016, 68% of adults in the US used Facebook, and 74% of adult Facebook users use Facebook *every day* (Smith and Anderson, 2018), providing both an audience and a distribution platform for online content. Software for managing online content has become nearly free, so the cost of setting up a website with the potential to reach the millions of daily Facebook users has fallen to nearly zero.

Equally important to the cultural climate is the regulatory environment. During the postwar era, the power of the Federal Communications Commission has been continually eroded. The 1987 repeal of the Fairness Doctrine, the overturn of the obscenity provisions of the 1996 Telecommunications Act and the 2000 repeal of the "personal attack" and "political editorial" rules were all decisions that lifted restrictions on what the media could do (Berry and Sobieraj, 2013). The First Amendment provides robust protections to freedom of the press, and attempts to regulate internet content have been met with extremely negative reactionsColeman (2014). The generational makeup of Congress makes it unlikely that key members understand the technical challenges of regulating the internet.

Another form of regulation impacts firm entry: relative regulatory licensing restrictions. In the real world, there are a finite number of entrepreneurs who plan to start a company. One important consideration is the explicit regulatory hurdles in each industry; industries with fewer licensing requirements are, all else equal, more attractive.

In order to charge money for barbering services in California, you must pass a written and practical examination. Before you may attempt to do so, you must have logged 1,500 hours of (unpaid) training barbering; barbering without a license is subject to a $1,000 fine (Department of Consumer Affairs and Cosmetology, 2016).

To set up a national media company based in California, you need to do exactly nothing; there are no licensing requirements and no potential fines. In 2013, Jestin Coler did just that: he established twenty or more media sites with next to zero editorial discretion, including the now-infamous Denver Guardian, which published stories with headlines like "FBI AGENT SUSPECTED IN HILLARY EMAIL LEAKS FOUND DEAD IN APPARENT MURDER-SUICIDE". Coler—who considers himself an entrepreneur—reportedly makes hundreds of thousands of dollars a year. His employees are all freelancers who work completely anonymously: none of them face any regulatory burden whatsoever (Sydell, 2016).

Although this is an extreme case, online media companies are almost always structured so that they get the majority of their content from freelancers, or at least from non-union workers.[2] This institutional setup means that web native companies have minimal long-term commitments, giving them close to zero exit costs.

[2]Employees at Gawker media, one of the most successful of the first wave of web native media companies, voted to unionize early in 2016. Although this is not necessarily related, the company was mere months away from being sued into bankruptcy.

0.4 Sunk Costs

Essential to operating a news media company is that consumers believe what you report. Developing a reputation for credible reporting is a pre-requisite for operating a serious national news outlet. This was not always the case; the establishment of professional journalistic practices was an explicit strategy of papers like the New York Times around the turn of the century in order to differentiate themselves from less reputable competitors (Ladd, 2011). With expanding competition from cable news and talk radio, however, it became clear that these high standards were no longer essential.

There are still a large number of discerning news consumers who only consume content from reputable outlets; this product differentiation means that the industry is not a perfectly contestable market. But the august reputations of legacy news outlets are simply not important to many consumers; as I argue below, this reputation has actually become a liability for many (especially conservative) consumers who distrust established media outlets.

The theory motivating high-quality journalism was that the expensive and time-consuming process of developing a reputation was a sunk cost that traditional outlets hoped would scare off new entrants—when a newspaper goes out of business, there is no way to recoup these reputation-building costs—but the modern news industry does not require these sunk costs. Gentzkow and Shapiro (2008) argue that a diverse media environment leads to greater investment in high-quality investigative reporting because of market discipline, but they also claim that "this mechanism will only operate if firms value a reputation for reporting the truth."

The mature online news industry represents the culmination of this de-valuing of reputa-

tion. Even with cable news or radio, reputation matters because consumers have to decide to change the channel to a specific media outlet. The same process occurred on the early web, where people had to decide which websites to visit. On social media, each individual piece of content competes with every other piece of content as individual users decide which pieces of content to share.

Reputation still could matter, though, to convince a potential reader that the reporting in a story is credible. The fact of social recommendation provides an alternative source of credibility; experimental evidence suggests that consumers prefer to share content with more anonymous recommendations on Facebook (Messing and Westwood, 2012). Social recommendations from friends or family members provide even more legitimacy. Furthermore, people are less likely to fact-check information they encounter on social media (Jun, Meng and Johar, 2017).

Consider the rational case for information acquisition. The likelihood that following a news story will cause an individual to change their views sufficiently to change their vote choice and that this will in turn change the outcome of an election is minimal; if this is the only benefit to information acquisition, ignorance is rational (Downs, 1957).

An alternative strand of scholarship, most famously advanced by Campbell et al. (1960) and seeing renewed interest with Green, Palmquist and Schickler (2002), conceptualizes partisanship as a social identity (Iyengar, Sood and Lelkes, 2012; Mason, 2016).[3] Instead of being rational truth-seekers aiming to make the best possible vote choice, partisans are team players who aim to follow social cues about the correct views and arguments for members of their social group.

[3]For an excellent intellectual history of the group theory of voter behavior, see Chapter 2 of Achen and Bartels (2016).

This story comports with the explanations given in qualitative interviews reported in Berry and Sobieraj (2013). People who consume what Berry and Sobieraj call "outrage media" do so because they want to feel like a member of a morally righteous social circle, and to feel educated: they want to have talking points ready to go for the next time they have a political discussion. Because these people are likely to avoid political discussions with members of the opposing party, these talking points serve primarily to establish their legitimacy with co-partisans.

As a result, the current technological/cultural environment is ripe for market segregation, enabling some media companies to attract niche audiences while sinking minimal costs into investing in credibility.[4] The media to which people are exposed on social media is explicitly the media that people in their social networks think is important, and because social networks are homophilious, this means that content should be spread among individuals who share a social identity.

This is the explicit strategy of web native media outlets like Gawker, Vice, and especially Buzzfeed. Buzzfeed's strategy is to make content that people want to share. As a trivial example of how this works, consider the series of quizzes that purport to explain where the reader is from (a form of social identity):

- "Your Cheesecake Factory Order Will Tell Us Which State You're From"
- "We Can Guess Where You're From Based On Your Bagel Choices"
- "Can We Guess The State You Live In Based On Your Restaurant Choices?"

[4]These effects are likely to be asymmetric. Republicans are more distrustful of the mainstream media and far more unified by symbolic ideology (Grossmann and Hopkins, 2016). In the course of the 2016 election, Allcott and Gentzkow (2017) estimate that the average American remembered .92 pro-Trump and .23 pro-Clinton stories from a zero-credibility outlet.

- "We Know Where In America You Actually Live"
- "Can We Guess Where In The USA You Actually Live?"

Even these outlets, though, decided to develop a brand identity of some kind, in order to get the initial consumers necessary to begin spreading their content. They also focused on the traditionally coveted (and digitally active and proficient) demographic of young, educated consumers. The 2016 US Presidential election demonstrated that even this minimal level of reputation-building sunk cost may no longer be necessary.

News reports in the wake of the election focused on the problem of "Fake News"—false or wildly misleading online content peddled by unknown media outlets and spread via social media. The most notorious example is the Denver Guardian, discussed above. In addition to employing anonymous freelancers, its business model was explicitly to publish outrageous stories that would appeal to partisan biases(Sydell, 2016).

The crucial innovation of these sites is to spread news that is verifiably false. For this to work, consumers must be unwilling or unable to consult other media sources and learn that they have been deceived.

In controlled settings, studies have shown various degrees of success in correcting misinformation, with partisans being less likely to accept that an ideologically congruent belief is false (Bode and Vraga, 2015; Garrett, Nisbet and Lynch, 2013; Nyhan and Reifler, 2010; Nyhan et al., 2014). These studies do an excellent job of delineating the necessary and sufficient conditions for the correction of misinformation, but they cannot speak to how often those conditions obtain in different media consumption contexts.

An exception is Jun, Meng and Johar (2017), who find that people are less like to

fact-check in social situations, like "platforms that are inherently social (e.g., Facebook) or...features of online environments such as 'likes' or 'shares.'" People tend to "let their guard down" when consuming news obtained in social settings, such as social media.

There is a large literature on the "digital divide" between people with internet access (or social media accounts) and those without (Chadwick, 2006; Mossberger, Tolbert and McNeal, 2007). This focus on the technological aspect of internet use has sometimes been too focused on the digital access binary, rather than the fact that skills are unevenly distributed among internet users (DiMaggio, Hargittai et al., 2001).

Hargittai (2001) calls this the "second-level digital divide": the wide disparity in the accuracy and speed at which internet users can perform even a standard task like information retrieval. The OECD performs the most comprehensive research on adults' skills and finds evidence of a massive disparity in the skill sets of the digital elite (a category to which essentially all producers of online media belong) and the majority of internet users (Kankaraš et al., 2016). The OECD's survey instrument is designed to measure skills related to "problem solving in technology-rich environments" (PSTRE), which they define as "Ability to use digital technology, communication tools and networks to acquire and evaluate information, communicate with others, and perform practical tasks." This measure is not explicitly designed to capture adults' capacity to determine the authenticity of a piece of news content on Facebook, but the skills it measures are closely related.

The extent of this "second-level digital divide" bears emphasizing. According to results published in 2016, US adults fall into one of five skill levels, each defined by the most sophisticated computer task people at that level can complete:

- Can't use computers (26%)
- Can delete an email (14%)
- Can use "reply-all" to send an email to three people (29%)
- Can "find a sustainability-related document that was sent to you by John Smith in October last year" (26%)
- Can calculate "what percentage of the emails sent by John Smith last month were about sustainability" (5%) (Nielsen, 2016)

These examples are all related to email and not directly relevant to evaluating a news story on Facebook; they are presented merely to demonstrate the level of difficulty of the tasks the OECD uses to calculate its PSTRE measure. Most of the tasks have to do with extracting information from specifically curated (and simple) web pages.

These numbers from the OECD are aggregates of all "adults": people aged between 16 and 65. However, there is a massive heterogeneity across age cohorts: in the United States, 39% of people aged 25-34 scored in the top two categories, but only 20% of people aged 55-65 did. It is overwhelmingly likely that the level of computer skills among those over 65 is even lower than for those in this age range. Combining these data with those concerning social media use leads to a startling conclusion:

68% of adults (this figure includes people over 65) use Facebook, but only 60% of adults (ages 16 to 65) are able to reply-all to an email.

The combination of these low-skill social media users and the power of social recommendation are necessary conditions for the zero-credibility firms that operate in the Facebook

Media context. By buying Facebook ads promoting their articles or pages, they are able to get the initial exposure they need. This process can happen without the majority of news consumers (the "mainstream audience") ever being aware because of the capacity for these social media sites to sell unprecedentedly well-targeted ads. Facebook touts this ability: in marketing material for advertisers, it claims that "[w]ith our powerful audience selection tools, you can target the people who are right for your business" (Facebook, 2016).

Zero-credibility firms target precisely these social media users who are least able to intuit or ascertain the accuracy of their content. These are the people who have access to the internet and who use social media and yet have low levels of digital literacy. A strong proxy for digital literacy is age, which scholars have found to be strongly predictive of Fake News consumption. Using web tracking data during the 2016 election, Guess, Nyhan and Reifler (2017) find that"only 7% of people age 59 or younger consumed one or more pro-Trump Fake News articles compared with 19% among those age 60 or older." In the empirical results below, I demonstrate that older people have a significantly higher preference for clickbait, controlling for a suite of other demographic variables.

If media sites are able to convince these less digitally literate people to consume and then share this inflammatory but fictitious content, their audience expands *and* the fact of the social recommendation lends legitimacy to their content.

In this way, people with higher degrees of skepticism or with greater digital literacy are enticed to consume and share content that they might otherwise find questionable. They become aware that other people with whom they share their partisan (social) identity are consuming this information, and they thus have reason to read it themselves, even though the

media organization that produced it has sunk absolutely nothing investing in the credibility of their brand.

As the story continues to be read and shared, the audience grows exponentially but also changes qualitatively as incrementally more and more sophisticated/digitally literate news consumers find the story credible. Updating the concept of the "information cascade" at the center of Hamilton (2004)'s theory, I call this process a *credibility cascade.*

Credibility cascade: Social recommendation provides credibility to news stories as they spread along online networks, cascading through layers of increasingly sophisticated digital news consumers.

Once a story accumulates enough social recommendations, it acquires sufficient legitimacy that more traditional news outlets cover it. This process represents a parallel credibility cascade, as increasingly high reputation media outlets cover the story, based on the "reporting" done by less established outlets. These tendencies in the traditional media have been well documented by other scholars of media and politics.

Boydstun (2013) describes a media industry characterized by path dependency and institutional conservationism. If a particular story begins to attract media coverage, other media organizations are likely to cover it as well: it has demonstrated an ability to attract audience attention. Once a media company has devoted investigative resources to a story, there is tendency to follow up on related topics, meaning that media attention to a particular topic is likely to persist.[5]

[5]This finding comports with a wealth of sociological evidence on the tendency for similar organizations to develop isomorphic

0.5 Technology/Skills

The necessary technology to create news content in contemporary Clickbait Media is unpatentable and (other than lengthy, in-person investigative reporting) within the reach of any incipient media company. The only inputs are internet access, computer hardware and digitally literate employees. The former two have decreased dramatically in price and require no special level of technical know-how to acquire and use. There is very little to differentiate one news story from another when they are both shared on a Facebook Newsfeed; a link to the (fake) *Denver Guardian* contains a headline, photo and caption, as does a link to the *New York Times*.[6]

The biggest technological change is in the kind of training necessary to be a journalist. The ideal of the Journalism Schools founded by Joseph Pulitzer and William Randolph Hearst was to ensure that all journalists had rigorous training in journalistic ethics and practice—journalists were to be respected professionals, on a par with doctors and lawyers. There was never a bar exam or explicit certification process, but there was the expectation that everyone in a newsroom would–either through J-school or on-the-job mentorship–share a broad skill set and ethical standards.

The technological constraints of broadcast television and newsprint meant that the supply of J-school graduates and job openings for journalists were in a rough equilibrium. Employees with professional degrees are expensive, though, and web native companies have not found it necessary to require their entry-level news reporters to have such formal training. Indeed, each of these freelancers is competing directly with other freelancers, often working for

institutional setups that decrease diversity of output (DiMaggio and Powell, 1983).

[6]This was true at the time of the 2016 election. Since then, Facebook has tested several display and algorithmic changes to Newsfeed that were designed to ameliorate this problem.

multiple news outlets at the same time; they have little organizational loyalty, instead aiming to raise their own profile. This is precisely the mechanism for media bias proposed by Baron (2006):

"[J]ournalists may bias their stories if their career prospects can be advanced by being published on the front page. News organizations can control bias by restricting the discretion allowed to journalists, but granting discretion and tolerating bias can increase profits if it allows journalists to be hired at a lower wage."

This leveling of journalistic skills is enabled by the medium of electronic distribution. Editors of print newspapers or cable news segments are space constrained: the inclusion of one story necessarily implies the exclusion of another. This is not the case on the web, which means that each story need not be held to the same standard of quality (in addition to the decreased importance of brands discussed above).[7]

The story of this transition is best told through aggregate employment data. Figure 0.1 displays the striking decrease in employment by newspapers—from over 450,000 jobs in 1990 to under 200,000 in 2016—and the concomitant increase in jobs in "internet publishing and broadcasting." Neither trend has significantly changed trajectory in the eight years post-recession.

The opportunity cost of each online story is zero, the marginal cost of distribution is zero, and the marginal cost of creation corresponds to the value of the journalists' time. Clickbait Media has taken advantage of the fact that there is a surplus of potential employees with

[7] In addition to the three major conditions outlined in above, Baumol specified that a contestable market needed to be one in which products were undifferentiated—the classic economics example points out that each bushel of wheat is indistinguishable from each other bushel. In one sense, the online media industry is maximally differentiated, as each piece of content is unique. In practice, though, each piece of content is just another combination of "words on a screen". If any piece of content is particularly popular, scores of imitators from other media companies will create versions of it that are only trivially different; often, these outlets will write a more inflammatory headline and introductory paragraph followed by a block quote from the original publication. These pieces of content, then, are close to undifferentiated.

Figure 0.1: Newspaper Employment Declines, Internet Publishing Employment Soars

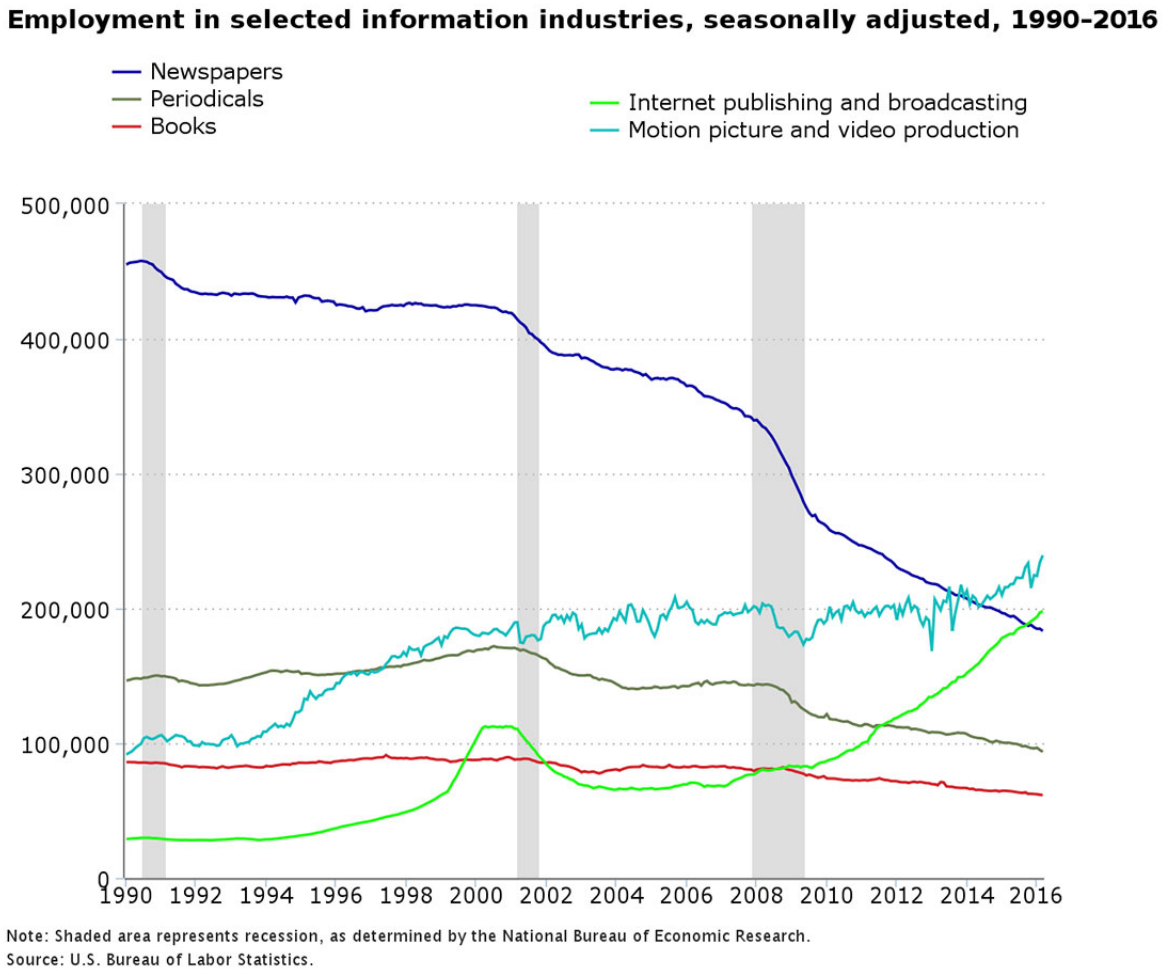
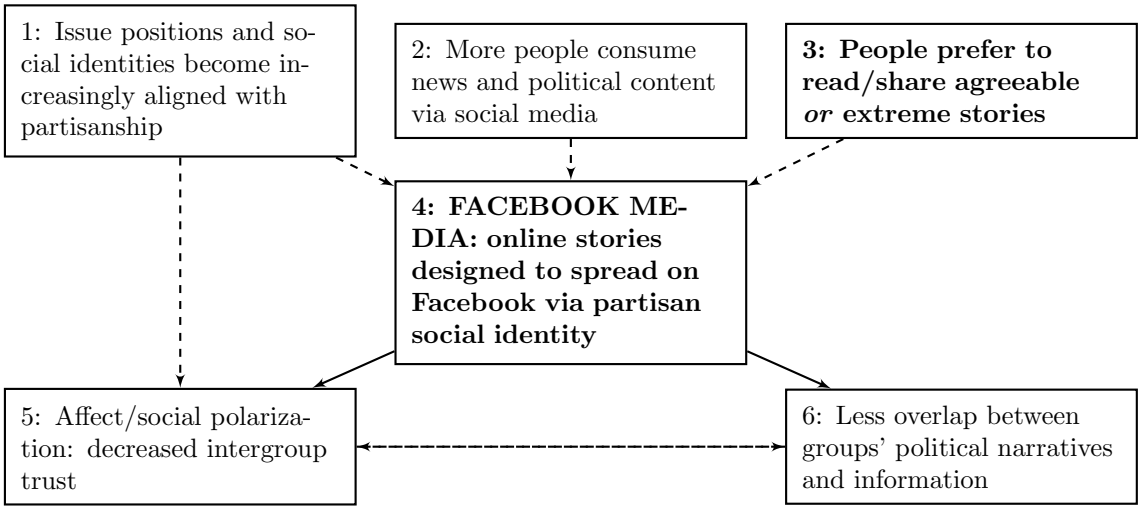


Figure 0.2: Overview of Concepts Related to Clickbait Media



the necessary skills: gathering information on the web, rapidly writing summaries of their findings, and possessing the social media/cultural savvy to promote their news content.

Indeed, it as if the US higher education system were designed to produce a surplus of graduates with precisely these skills.

0.6 Connecting Modern Media Trends and Outcomes

The theoretical argument is outlined in Figure 0.2. The bolded cells are the focus of this piece; the theoretical argument above has argued that point 4, Facebook Media, accurately describes the structure of on the online news industry. The solid lines are the causal pathways I test below, while the dotted lines indicate effects that I believe exist but which are less central to my analysis. I take points 1, 2, 5 and 6 as established in the literature.[8]

The experiments in Chapter 3 aim to establish point 4 with the clickbait headline format

[8]Point 6 is currently controversial; the existence of "filter bubbles" or "echo chambers" is being hotly debated (Bakshy, Messing and Adamic, 2015; Barberá, Jost, Nagler, Tucker and Bonneau, 2015; Flaxman, Goel and Rao, 2013). The most recent (and in my reading, best) analysis of the phenomenon finds that these bubbles exist, but only for a small portion of especially partisan citizens (Guess, Nyhan and Reifler, 2017). Point 6 should thus be taken as applying to a small but politically influential group.

most relevant to partisan sharing: *emotional clickbait*. To operationalize this concept, I add emotional cues to the beginning of partisan headlines ("This will make your blood boil:..."; "Republicans are furious:...) and estimate the treatment effect of being exposed to *emotional clickbait* on points 5 and 6.

One important caveat to my general intellectual project here is that the importance of a single privately-held company is a severe limitation to the generalizability of the theory I have developed: Facebook could, tomorrow, radically change its relationship with media companies, rendering my theory obsolete.

This problem is more general, extending to all social science research that takes the internet seriously. As David Karpf argues in his excellent article on *Social Science Research Methods in Internet Time*, "(1) The rate at which the Internet is both diffusing through society and developing new capacities is unprecedented. (2) Many of our most robust research methods are based upon *ceteris paribus* assumptions that do not hold in the online environment."(Karpf, 2012*b*). The rate of change of the object under study—the way that citizens use the internet, for example—is high and rising. Social scientists are constrained by the rhythms of academic publishing, limiting our ability to accumulate knowledge about the internet according to the scientific method.

Take, for example, Bond et al. (2017), which provides subgroup breakdowns of the results in the classic Bond et al. (2012) study of social influence on Facebook. At the time of writing, the original study has over 1,200 citations, and presents two main findings: (1) social messages (in which the faces of subjects' friends who voted were included along with information about how to vote) caused a significant increase in validated voting; (2) non-

social, informational messages had a precisely estimated null effect on validated voting.

The new study was published in 2017, using data from the original experiment conducted on Facebook in 2010. One of the findings in the new study is that the effect of social recommendation is much larger for older people than for younger people. To argue that this finding is relevant to the theory of credibility cascades I developed above requires the following assumptions:

- The older Facebook users in the 2010 study are representative of the older Facebook users who consumed zero-credibility news in 2016.
- The Facebook interface has not changed between 2010 and 2016.
- The specific estimand in the study is generalizable to other estimands.

These assumptions are, necessarily, false. What are the implications for applying the finding from 2010 to the 2016 election? Answering that question would require conducting more studies, which would in turn become potentially outdated as quickly as they were published. Indeed, the 2010 experiment was replicated in the 2012 Presidential election by Jones et al. (2017). The cultural-technological environment had changed since 2010, and this election was Presidential instead of merely Congressional; the results were *very* different. The point estimate of the identical treatment on validated voting was .39 in 2010 and .17 in 2012; the latter was not significant without the addition of control variables. How much of this difference is due to the changing cultural-technological environment and how much due to the electoral context?

These problems are not fatal, but social scientists who study human behavior on the internet should be aware of them and take steps to ameliorate them. The biggest changes

that need to be made are structural, at the level of disciplinary standards, and progress has been made: the movement to make data and code publicly available, the trend of posting working papers, and the creation of new journals or journal formats that encourage short publications all help us produce knowledge faster.

Social scientists have learned to be concerned with the issue of external validity when taking the results of a study in a given context and applying them more generally; there is an intrinsic trade-off between gaining ironclad knowledge that is deeply situated in a given context and creating more generalizable knowledge.

I hope that we can also learn to take this problem of *temporal validity* seriously. The best work should strive to be internally, externally and temporally valid.

Chapter 1

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

1.1 Introduction

The explicit expression of hostile prejudice is no longer acceptable in mainstream US society. This is evidence for changing social norms, though these new norms are not as well-established in some communities, especially on the internet. The rise of online social interaction has brought with it new opportunities for individuals to express their prejudices and engage in verbal harassment.

This behavior has implications for both the perpetrators and their victims. Minorities and other vulnerable populations are frequently the subject of online harassment on social media sites, often in response to expressing views that harassers disagree with (Kennedy and Taylor, 2010; Mantilla, 2013). They are likely to become more anxious for their safety, more

fearful of crime and less likely to express themselves publicly (Henson, Reyns and Fisher, 2013), systematically de-mobilizing the populations who tend to be victimized (Hinduja and Patchin, 2007). Engaging in harassment of non-whites also fuels ethnocentrism among whites, which has been shown to affect how whites feel about political topics like healthcare and immigration (Banks, 2014, 2016), and to affect voting outcomes (Kam and Kinder, 2012).

Severe online harassment takes the form of explicit threats or the posting of personal information, forcing targets to modify their behavior out of fear for their immediate safety. Although all harassment can contribute to a toxic online community, this paper is specifically about racist harassment of white men against blacks.

There have been many efforts to reduce online harassment on the part of online forums for social interaction, as well as by brick-and-mortar institutions like schools, universities and government agencies. They tend to involve blanket bans on certain behaviors, enforced either through the public promotion of norms or individual sanctions for clear violations enforced by moderators. A comprehensive review of the literature on prejudice reduction and harassment prevention (Paluck and Green, 2009) finds that very little of the research in this area is causally well-identified, and calls for more experimental research. I conducted a novel randomized field experiment that is able to measure the causal effect of specific interventions on the real-world harassing behavior of Twitter users, continuously and over time.

I searched for tweets containing a powerful racial slur ("n****r") to identify harassers with public Twitter accounts, and I assigned each subject to the control or to one of four treatment conditions. Using Twitter accounts that I controlled ("bots"), I tweeted at the subjects to tell them that their behavior was unacceptable. I varied two aspects of the bots, resulting in a 2x2 experimental design: the first dimension of variation was the identity of the bot, to test the finding from Social Identity Theory that sanctioning by members of a person's in-group is more effective (Tajfel and Turner, 1979). The second variation was in

the number of followers the bot had. This tests the "influentials hypothesis," that influential individuals are crucial for driving changes in norms of behavior in society (Aral and Walker, 2012).

I also expected to find heterogeneous treatment effects in the degree of anonymity of the subjects. Based on findings from related online contexts (Omernick and Sood, 2013), my hypothesis was that more anonymous individuals would be less likely to respond to the treatment. An alternative hypothesis, based on the Social Identity model of Deindividuation Effects (SIDE), would be that more anonymous individuals would actually be *more* susceptible to this normative pressure (Postmes et al., 2001).

I find support for the hypothesis that the same message had disparate impact based on in-group identity (here, race), with messages sent by white men causing the largest reduction in offensive behavior among a subject pool of white men.[1] However, this effect was only found among messages sent by accounts that had a high number of Twitter followers. This effect persisted for a full month after the application of the treatment. This finding concords with my hypothesis that the largest treatment effect would be that of receiving a message from a high-status white man. However, the effect of the followers treatment and the group identity treatment were *multiplicative*, rather than additive, as none of the other treatment conditions caused a significant behavioral change.

The results varied by the degree of anonymity of the subjects. The main effect was substantively similar among the anonymous subgroup. Among the subjects who provided some amount of identifying information, though, the reduction disappeared, and there was actually an *increase* in racist harassment among the subjects who received a message sent by a black bot with few followers. This finding was contrary to my hypothesis, and lends support to the role of anonymity in the SIDE model in this context.

The net effect of all of the treatments in this study was to reduce the rate of racist harassment. Overall, in the one month post-treatment collection period, my intervention

[1] All hypotheses were pre-registered at EGAP.org prior to any data collection.

caused the 50 subjects in the most effective treatment condition to tweet the word "n****r" an estimated 186 fewer times in the month after treatment.

1.2 Reducing Manifestations of Prejudice

Racism, which is a necessary component of the racist harassment studied here, is a form of prejudice, which Dovidio and Gaertner (1999) define as an "unfair negative attitude toward a social group or a member of that group," and Crandall, Eshleman and O'Brien (2002) define as "a negative evaluation of a group or of an individual on the basis of group membership." This paper makes the assumption that directing the word "n****r" at another person constitutes racist harassment, regardless of how justified the user believes their prejudice to be.

Beginning with Allport (1954)'s influential work on prejudice, the subject has been well-studied in psychology. Allport's "contact hypothesis"–that mere contact between different groups helps to reduce prejudice that each holds towards the other–has proven difficult to verify causally. A comprehensive review finds only mild support for the contact hypothesis (Pettigrew and Tropp, 2006), and others note that the subject makes isolating causation difficult (Binder et al., 2009).

A more promising approach for analyzing the formation and reduction of prejudices has to do with social norms. Group norm theory holds that "social norms [including prejudices] are formed in group situations and subsequently serve as standards for the individual's perception and judgment when he is not in the group situation" (Sherif and Sherif, 1953). Attitudes towards out-groups are a particularly important set of group norms, and prejudice towards out-groups can be a strong signal of in-group membership (Brewer, 1999).

Recent experiments have aimed to test the role of group norms in prejudice formation. Prejudiced attitudes can be reduced (in the short term) by priming less prejudiced social identities; by increasing individual salience vis-a-vis group membership; and by using a

confederate to challenge people's understanding of group norms (Blanchard et al., 1994; Dovidio and Gaertner, 1999; Plant and Devine, 1998). These papers, and others in the literature, suffer from a limitation common to experiments run with convenience samples: they cannot track either long-term or non-lab manifestations of prejudice. Two exceptions to the former problem are Stangor, Sechrist and Jost (2001), who show that providing consensus information about in-group norms of prejudiced attitudes can affect survey responses a week later; and Zitek and Hebl (2007), who find that social pressure is more effective at changing prejudiced attitudes if the norms are less clear (eg prejudice against obese people) up to a month after the experiment. By studying the behavior of people on Twitter, my approach is able to capture a continuous measure of prejudice reduction over time and in a naturalistic setting.

Although openly harassing people based on their race is not as common now as it once was, online racist harassment is an increasingly large problem. Studies of Computer Mediated Communication (CMC) have some insight as to why: CMC tends to result in less success in applying normative pressure (Bordia, 1997; Kiesler, Siegel and McGuire, 1984; Walther, 1996).

The primary mechanism used to explain the differences in CMC over the internet has been postulated to be *deindividuation*: people become immersed in the medium of discussion and lose a sense of self-awareness. This mechanism is best explained by the Social Identity model of Deindividuation Effects (SIDE), in which the depressed sense of one's personal identity is supplanted by an increased sense of one's social identity (Lea and Spears, 1991; Reicher, Spears and Postmes, 1995).

The anonymity enabled by CMC also leads to more racist harassment online. As Moor (2007) describes anonymous online communities, "people are relatively indistinguishable and their memberships of online discussion groups are far more salient than their personal identities." In communicating online, there are fewer dimensions on which people can identify with a group; speech norms are central.

Prejudiced harassment against out-groups has been used to signal in-group loyalty in the physical world, and it serves the same purpose in online communities. Engaging in prejudiced harassment against out-groups–in this case, blacks–primes ethnocentrism and changes the salience of particular political issues like healthcare (Banks, 2014) and immigration (Banks, 2016). There is also evidence that the expression of prejudiced views online has implications for vote choice, with the most prominent example being the 2008 presidential election. Increased belief in racial stereotypes decreased Barack Obama's vote total (Kam and Kinder, 2012; Piston, 2010).

But SIDE also suggests an avenue for reducing online racist harassment: individuals' social identities are actually composed of several overlapping identities. It follows that the influence of specific online communities with norms of online harassment can be diminished by appealing to their other, offline identities. Rather than leading to increased self-regulation and decreased responsiveness to normative pressure, as in classical models of deindividuation, SIDE posits that deindividuation–when enabled by anonymity–should lead to *increased* response to normative pressure (Postmes et al., 2001).

Still, as Paluck and Green (2009)'s summary of the literature points out, there has been little research done in the field of prejudice reduction using randomized experiments outside of the laboratory. This paper attempts to address this lacuna. It also represents, with Coppock, Guess and Ternovski (2015), one of the first randomized control experiments to be conducted entirely on Twitter.

The crucial advantage of this experimental design is that I could measure real behavior continuously for months. In order to quantify this behavior, I operationalized racist online harassment in the form of the use of the word "n****r." This slur is the most substantively important vehicle for racist harassment, and filtering on its use was the fastest way to collect a sample of genuine harassers. I acquired this data by scraping the Twitter history of each subject before and after being treated.

There is a sizable body of research that indicates that attempts to reduce prejudiced

behavior are more effective when made by members of the in-group (Gulker, Mark and Monteith, 2013; Rasinski and Czopp, 2010). There is also evidence that prejudice-reducing efforts made by higher-status individuals are more effective, although the exact definition of "high status" depends on the context. Paluck, Shepherd and Aronow (2016) find this to be the case when the high status individuals are "social referents" (who other students look to) in a high school, and Shepherd and Paluck (2015*b*) call highly-connected male high schoolers "high status." Aral and Walker (2012) observe differences in peer influence with a large-scale study of Facebook users, finding that influence varies with marital status and gender. In all of these contexts, the theoretical expectation is that "high status" individuals have a greater capacity to define group norms, and that observers are more likely to mimic their behavior to try and fit in with their group.

Because Twitter is a semi-anonymous environment, I draw from both the SIDE literature about group norm promotion and the research on highly influential social referents to motivate my hypotheses and related experimental manipulations. Specifically, I varied the identity of the bots applying the treatment. They were either In-group (white men) or Out-group (black men), and either had many followers or few followers. Based on the findings discussed above, my hypothesis was that the largest treatment effect would be from In-group/High Followers bots and that the smallest treatment effect would be from Out-group/Low Followers bots. I hypothesized that the other two treatment conditions would have medium-sized effects:

Hypothesis 1 *The ranking of the magnitudes of the decrease in harassment will be:* *In-group/High Followers* $> \begin{smallmatrix} In-group/Low\ Followers \\ Out-group/High\ Followers \end{smallmatrix} >$ *Out-group/Low Followers.*

Previously, the degree of anonymity allowed in an online community has been shown to affect the prevalence of online harassment, with more anonymity being associated with more harassment (Hosseinmardi et al., 2014; Omernick and Sood, 2013). Twitter allows users to be anonymous to the extent that their accounts can be entirely divorced from their real-life

persona, but many users choose to provide identifying information like that which identifies my bots.

To create an anonymity score, I examined several aspects of each subject's profile: whether they had a Profile Picture of themselves[2] and whether a given name was present in their username or handle. I used these to create a categorical Anonymity Score that ranged from 2 (most anonymous) to 0 (least anonymous).

The above findings about online communication suggest that greater anonymity is associated with more harassment and lower-quality communication, but SIDE theory implies that norm promotion should be stronger in anonymous contexts. The idea is that individuals make less of a distinction between themselves and other members of their group, and are thus more likely to follow group norms than their own idiosyncratic preferences (Postmes et al., 2001).

Neither of these strains of research have direct implications for my experimental design. Here, anonymity is a *self-selected covariate* of each subject, rather than a global characteristic. My expectation was that subjects who elected to share less personal information would be less invested in their online communities, and thus less likely to care about group norms. My findings show that this expectation was mistaken. The opposite turned out to be the case, with the expected treatment effects found only among the anonymous subjects.

Hypothesis 2 *The magnitude of the decrease in harassment will negatively covary with the subject's Anonymity Score.*

1.3 Experimental Design

Among the most challenging aspects of studying mass behavior on Twitter is the selection of a meaningful sample of Twitter users. In order to ensure that efforts to reduce racist harassment could be measured, it was essential to have a sample of users who engaged in

[2]Whether a picture is actually of the subject was impossible to verify perfectly; I included any picture that clearly showed the face of a person who I did not recognize.

racist harassment in the first place.

There is a large and growing literature on the automatic detection of online harassment (Chen et al., 2012; Yin et al., 2009). The task of discerning genuine harassment from heated argumentation or sarcastic joking is challenging, but the presence of *prima facie* offensive language makes it far easier. In fact, in corpuses that contain enough strongly offensive language, a simple dictionary of strongly offensive terms outperforms even sophisticated classifiers. The dictionary approach also has the advantage of being rapidly implementable at scale.

The detection of second-person pronouns, to determine at whom the profanity is directed, is a large and easy improvement on naive profanity detection, and the structure of Twitter use makes this kind of analysis straightforward: tweets that begin with an "@[username]" are explicitly targeted at the recipient. To further refine the search for *racist* online harassment, I created a sample of individuals who tweeted a racial slur ("n****r") at another account.[3] In the racial context of the United States, this term is almost certainly the most intrinsically offensive, and people who use it thus represent a "hard case" for this experimental design– there is no doubt that these people are aware that directly tweeting this term at another person constitutes harassment.

Using the streamR package for R, I scraped the user information (including the most recent 1,000 tweets) of anyone who tweeted the word "n****r" at another user. For each of these users, I applied a simple dictionary method to calculate the average number of offensive words per tweet in the text of those tweets to generate an offensiveness score for that user. As Sood, Antin and Churchill (2012) point out, the problem of selecting a list of "offensive" words is challenging, and some previous efforts have used arbitrary external dictionaries.[4]

[3] As is recorded in my Pre-Analyis Plan (registered at EGAP), I had originally intended to perform two similar experiments: one on racist harassment, and the other on misogynist harassment. However, my method was insufficient for generating a large enough sample of misogynist users. For any misogynist slur I tried to use as my search term (bitch, whore, slut), there were far too many people using it as a term of endearment for their friends for me to filter through and find the actual harassment. I plan on figuring out a way to crowdsource this process of manually discerning genuine harassment, but for now, the misogynist harassment experiment is unfeasible. The Pre-Analysis Plan also intended to test two hypotheses about spillover effects on the subjects' networks, but this has thus far proven technically intractable.

[4] Chen et al. (2012), for example, emulates Xu and Zhu (2010) and takes a list of terms from the website www.noswearing.com.

To avoid false positives, I used a much shorter list of swear words and slurs.[5]

I discarded users whose offensiveness score fell below a certain threshold and who were thus not regularly offensive. To determine what this "regularly offensive" threshold should be, I randomly sampled 450 Twitter users whose accounts were at least 6 months old.[6] I calculated the offensiveness score for these users' most recent 400 tweets and set the threshold for inclusion in the experimental sample at the 75th percentile of offensiveness. Substantively, this meant that at least 3% of their tweets had to include an offensive term.

This addressed many problems that could arise from the use of jokes or sarcasm: a dictionary method like searching for ethnic slurs cannot capture any information about the tone of a tweet, but leveraging more data and richer contextual information makes misclassification less likely.[7]

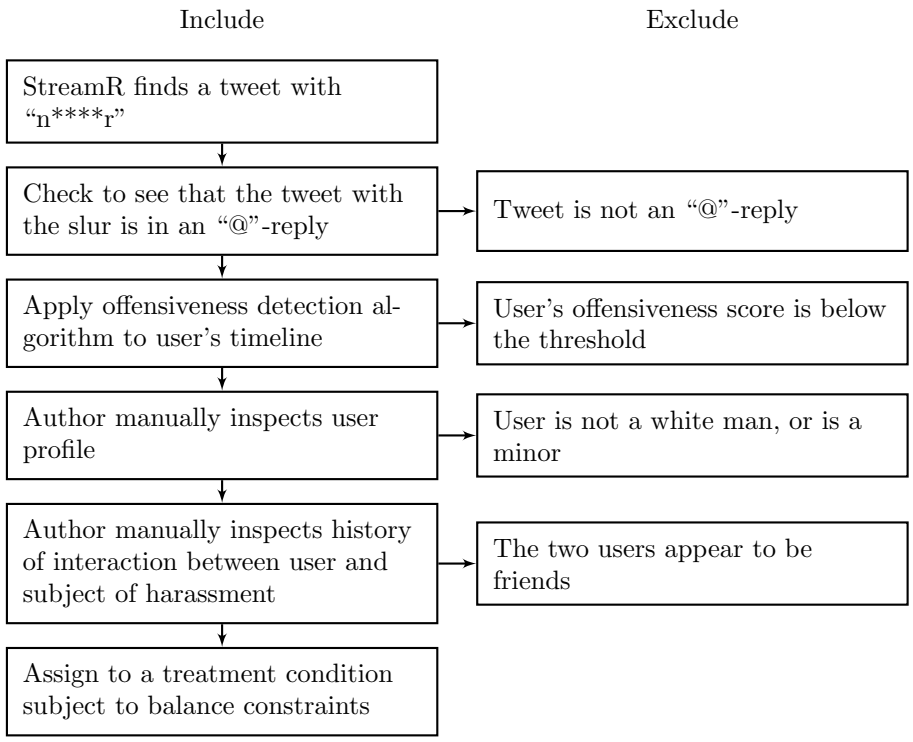
There were several other restrictions I placed on the sample of users. Because they are the largest and most politically salient demographic engaging in racist online harassment of blacks, I only included subjects who were white men. This ensured that the in-groups of interest (gender and race) didn't vary among the subjects, and thus that the treatments were the same. This additional control was essential, given the power of the study. I also included anonymous users because there were a large number of such accounts engaging in prejudiced harassment and I had different theoretical expectations about how they would respond to treatment. I recorded the degree of anonymity on a categorical scale from 0 to 2 based on if they included their real name and/or a picture of themselves. To the extent possible, I also excluded minors from the sample. Most users did not provide their exact age, but I removed from the sample any user who gave an indication of being underage or who mentioned high school.

[5]For a full list of terms, see the Online Appendix.

[6]Each Twitter account is assigned a unique numerical User ID based on when they signed up; newer accounts have higher ID's. Not all of the numbers correspond to extant or frequently used accounts, so if I randomly picked one of those numbers, I generated a new random number.

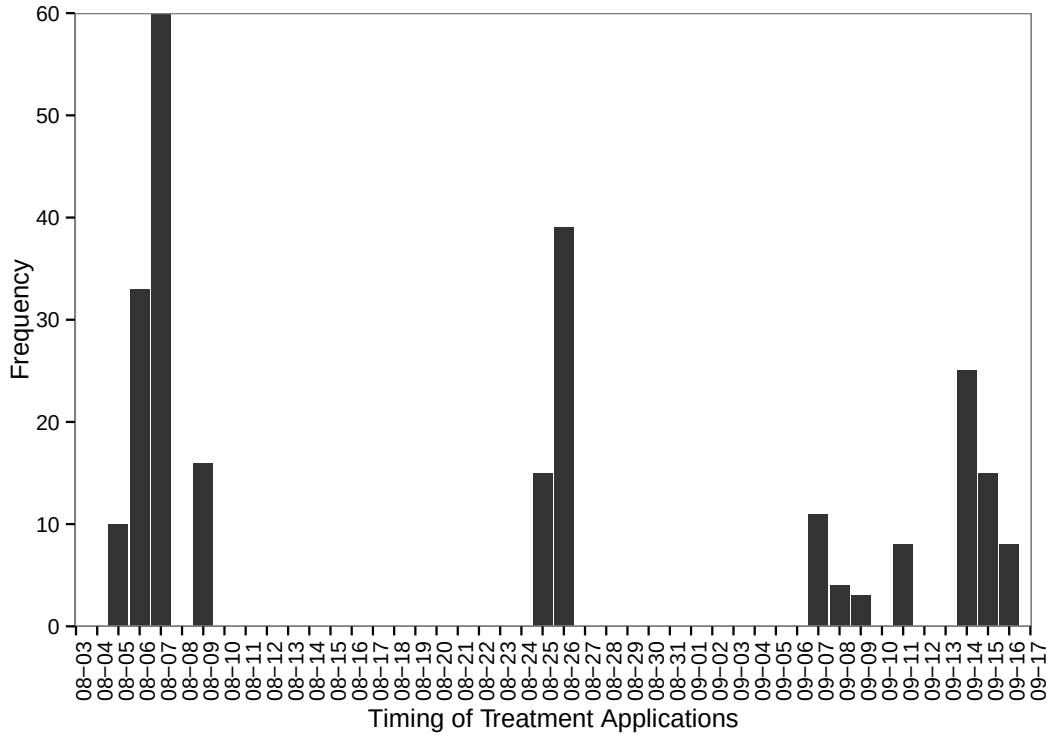
[7]Still, there are many people who believe that they're "joking" when they call a friend a slur. While this is still objectionable behavior, it is different from the kind of targeted prejudiced harassment that is of interest in this paper, so I excluded from the sample any users who appeared to be friends who did not find the slur they were using offensive. This process is inherently subjective, but it usually entailed the users with a long back-and-forth, with slurs interspersed with more obviously friendly terms.

Figure 1.1: Sample Selection Process



This flowchart depicts the decision process by which potential subjects were discovered, vetted and ultimately included or excluded.

Figure 1.2: Timing of the Experiment in the Field



The number of subjects added to the sample each day is plotted on the y-axis. Each treatment was applied within 24 hours of the subject tweeting a racial slur. There were potential subjects tweeting every day, but I was only actively searching on the days indicated. All dates 2015.

Because the subjects in this experiment were drawn from a specific subsection of the overall population, the criteria for inclusion discussed above are fundamental. Figure 2.2 provides a visual overview of the sampling procedure.

Table 1.1: Experimental Design and Hypothesized Effect Sizes

	In-group	Out-group
Low followers	Medium effect	Small effect
High followers	Large effect	Medium effect

After I verified that a user met all of the criteria for inclusion, I assigned him to one of the treatment conditions or the control condition, subject to balance constraints.[8] Because this process was time-consuming, and there were a fixed number of potential subjects who met these criteria tweeting at a given time, the subject discovery and vetting took place in several periods. The first wave of subjects was collected from August 5th to August 7th, 2015; the second wave from August 25th to August 26th; the third wave from September 7th to September 11th; and the last wave from September 14th to September 16th. See Figure 1.2 for a visual summary.[9] The crucial advantage of this real-time detection was that the time that elapsed between when a user tweeted the slur and when he received the treatment was under 24 hours, adding to the realism of the treatment.

The actual application of the treatment was straightforward. Depending on which condition the subject was assigned to, I rotated through the bots in that condition and tweeted the message:

```
"@[subject] Hey man, just remember that there are real people who are
hurt when you harass them with that kind of language"
```

Because this was an "@"-reply, it was only visible to anyone who clicked on the harassing tweet, and to the subject himself.

The four experimental conditions are summarized in Table 1.1. I varied the race of the bots in order to test the findings in Rasinski and Czopp (2010) and Gulker, Mark and Monteith (2013) that in-group sanctioning is more effective than out-group sanctioning: in

[8]Throughout the assignment process, I matched subjects in each treatment group on their (0 to 2) Anonymity Score. They were otherwise randomly assigned.

[9]This process was approved by NYU's Institutional Review Board. These subjects had not given their informed consent to participate in this experiment, but the intervention I applied falls within the "normal expectations" of their user experience on Twitter. The subjects were not debriefed. The benefits to their debriefing would not outweigh the risks to me, the researcher, in providing my personal information to a group of people with a demonstrated propensity for online harassment.

this case, that the effect of a tweet from a white Twitter user would be greater than one from a black Twitter user. The number of followers a Twitter user has is indicative of how influential they are, at least within the context of Twitter, so I varied that quantity to test the finding in Shepherd and Paluck (2015*b*) and Paluck, Shepherd and Aronow (2016) that sanctioning by high-status individuals is more effective than that by low-status individuals.

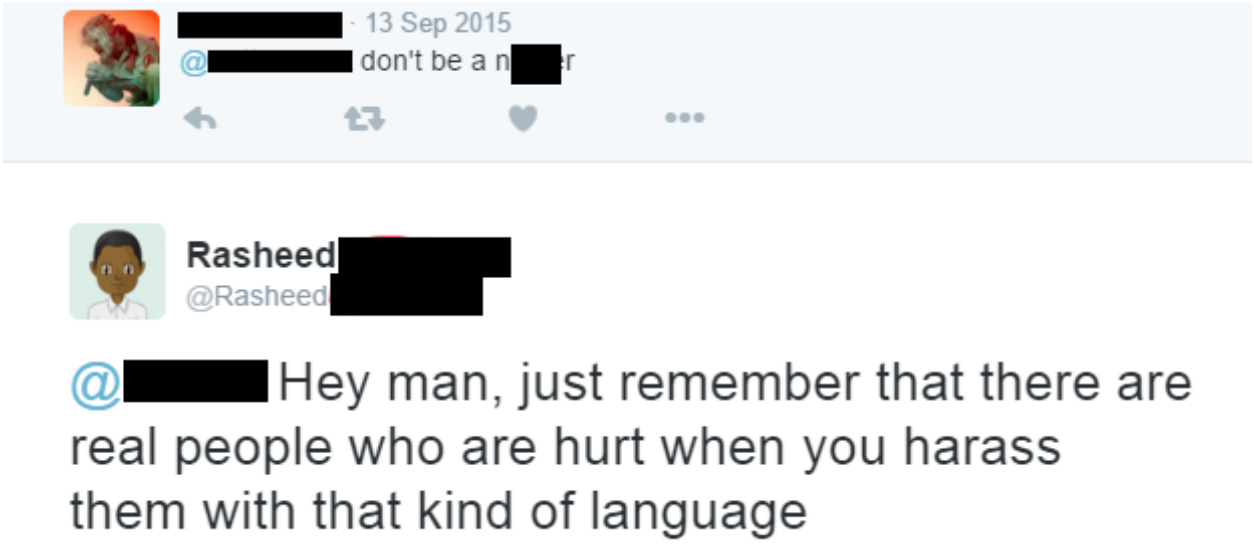
For example, users assigned to the Out-group/Low Followers condition were sent a message like the one seen in Figure 1.3(a), sent by bot @Rasheed[XXXXXX].[10] After the subject received the treatment, he got a "notification" from Twitter, which caused him to be exposed to the treatment tweet. Because being admonished by a stranger is an uncommon (though far from unknown) experience, the subject was inclined to click on the bots' account; if he did, he saw the bot's profile page, Figure 1.3(b). @Greg[XXXXXXX] was a bot in the In-group/Low Status condition. This allowed the subject to clearly determine the race and gender of his admonisher, and to see how many followers the account had (in this case, 2). I could not, however, directly measure this behavior, and it is possible that some subjects did not click on the bot's profile. If that were the case, they would still have noticed the bot's race from the profile picture and username, but they would not have seen the number of followers. This would bias the effect of the Followers treatment downward.

As the two bots shown in Figure 1.3 illustrate, the variation in the bot identity was accomplished by changing the number of followers, the skin color of the profile picture, username, and full name. To vary the number of followers, I bought followers for some accounts and not others (Stringhini et al., 2012). In the low-follower condition, the bots had between 0 and 10 followers (some of the bots were followed by other Twitter users, most of them spam accounts). In the high-follower condition, they had between 500 and 550 followers.

When generating the bots, I chose handles that consisted of first and last names that were identifiably male and white or black, following Bertrand and Mullainathan (2003). Because

[10] I avoid providing the entire username of the bot to protect my subjects' anonymity.

Figure 1.3: Treatments



(a) The treatment–black bot

(b) The bot applying the treatment–white bot

all of these handles were already taken (and Twitter requires that each account have a unique handles), I added random numbers to generate unique handles. The usernames were the first and last name used in the handle without the numbers; usernames do not need to be unique.

The most important aspect of the bots' profile was their profile picture. It was the first thing the subject saw, and was also the largest potential source of bias. In order to maximize the amount of control I had over the treatment, I used cartoon avatars for the profile pictures. If I had used real photos, there would exist the possibility that the particular people pictured varied on some important dimension other than race. This practice does not detract from the verisimilitude of the bot–using cartoon avatars on Twitter is not uncommon. I gave each bot the same facial features and the same professional-looking attire; the only thing I varied was the skin color, using a similar technique to Chhibber and Sekhon (2014).[11]

In order to ensure that the actual treatment experienced by the subject was maximally similar to the "real life" experience of being sanctioned by a stranger on Twitter, it was essential that the subject be unaware that my bot was in fact a bot. If the subject suspected that the bot was not the authentic online manifestation of a concerned citizen, the effects of norm promotion would be attenuated and the measured treatment effect would be a conservative estimate of the true treatment effect. One possible source of skepticism was that the followers I bought were not high-quality followers, in that they were obviously not real accounts; however, having fake or "spam" Twitter followers is not uncommon.

The history of tweets by the bot represented the most serious problem for verisimilitude. Under the "Tweets" tab displayed in Figure 1.3(b), there needed to be a plausible history of tweets to convey that this was a real, active user. To that end, I had the bot tweet from a list of personal but innocuous statements ("Strawberry season is in full swing, and I'm loving it") and retweeted a number of generic news articles. However, in the default profile display, tweets that are directed "@" another user are not visible. If the subject clicked on

[11] It is possible that a stronger racial treatment effect might have obtained if I also changed the facial features of the black bots to be more afrocentric, the effect of which Weaver (2012) finds to be approximately as large as changing skin color on voting outcomes.

Table 1.2: Attrition Rates

	Control	In-group Low	Out-group Low	In-group High	Out-group High
Baseline # of subjects	51	49	44	50	48
# with > 1 Post-treatment tweets	49	47	42	47	47
# with > 25 Post-treatment tweets	43	42	38	41	46
Attrition %, < 25 Post-treatment tweets	16%	14%	14%	18%	4%

The number of subjects who tweeted more than 1 or 25 times after the application of the treatment.

the "Tweets & replies" tab, they became visible, but my innocuous tweets were interspersed so that the treatment tweets represent less than half of the bot's overall tweets. As a result, only three of the 242 subjects responded to accuse my bots of being bots.

1.4 Results

The primary outcome of interest was the change in the subjects' levels of offensiveness in the four different treatment arms, relative to the control group. However, I could not collect a full two month's worth of tweets for some of the subjects, for one of three reasons: at some point after the treatment, the subject could have made his account private, or he could have deleted his account, or the account could have been banned by Twitter. The first only happened to three accounts out of the 242 in the sample,[12] but I could not distinguish between the last two.[13] Table 2.2 presents the attrition rates among the different treatment arms in the sample. The average attrition (defined as subjects who dropped out of the sample before tweeting at least 25 times after the treatment) among the four treatment conditions was 16%, compared to 13% among the control subjects, an insignificant difference ($p = .58$).[14]

Despite this insignificance, performing the analysis only on the subjects who remained in

[12]Initially, I assigned 243 subjects to one of the 4 treatment arms or to the control group. However, the rate of tweeting of one of these subjects was too infrequent for me to be able to calculate a meaningful pre-treatment rate of offensive language use, and I excluded him.

[13]I contacted Twitter to see if they could provide me with this information, but they were not forthcoming.

[14]Note, though, that the Out-group/High Followers condition saw much lower attrition than the other treatment conditions. I have no explanation for why this is the case, and in fact my ex ante expectation was that, to the extent that attrition was positively correlated with any treatment condition, it would have been higher among the High Followers conditions.

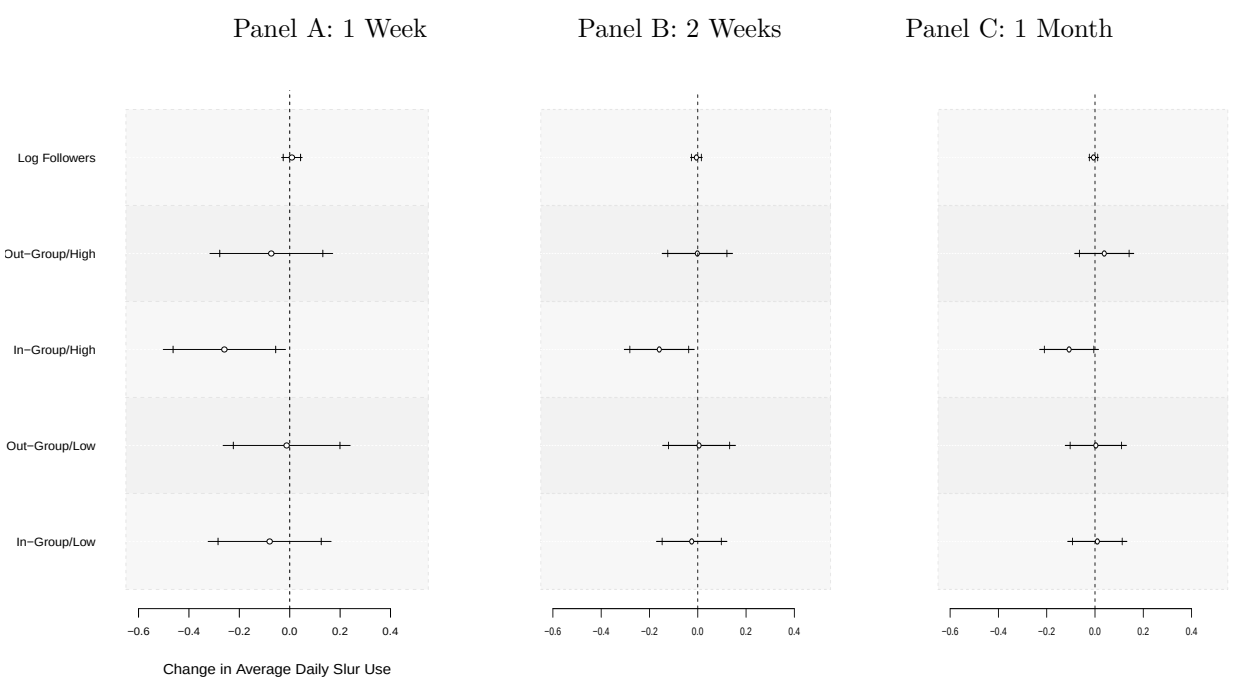
the sample could introduce post-treatment bias. It is preferable to include all of the subjects, but this requires an assumption about the behavior of the subjects for whom I had missing data. I made the following assumption: for each of these observations with missing post-treatment data, I treated their post-treatment rate of racist language as zero. The subjects who were no longer tweeting publicly had ceased to engage in online harassment.[15]

The results support H_1. In Figure 1.4, Panel A shows the effect of the different treatment arms on the absolute daily use of the word "n****r" over the week after the treatment.[16] Panel B expands the time period to two weeks, and Panel C expands it to one month. Each panel shows the result of an OLS regression in which the dependent variable is the absolute number of instances of racist language during that time period divided by the number of days in that time period. Each regression controls for the subjects' log number of followers, displayed in the first and second rows. Each regression also controls for the average rate of the subjects' use of that offensive term in the two months prior to the treatment. The four treatment arms each represent the comparison between that arm and the control group, and each treatment effect is displayed in one of the bottom four rows.

[15] A more conservative and less substantively accurate assumption is to treat these observations as having a post-treatment rate of racist language equal to their pre-treatment rate of racist language use. Appendix A presents the results with this alternate assumption. The results are substantively similar, although the point estimates are slightly smaller.

[16] I have selected my sample based on their use of this slur. Expanding the dependent variable to include other anti-black language does not substantively change the results, primarily because the use of other anti-black slurs is uncommon among this subject pool.

Figure 1.4: Reduction in Racist Slurs, Full Sample (*N=242*)



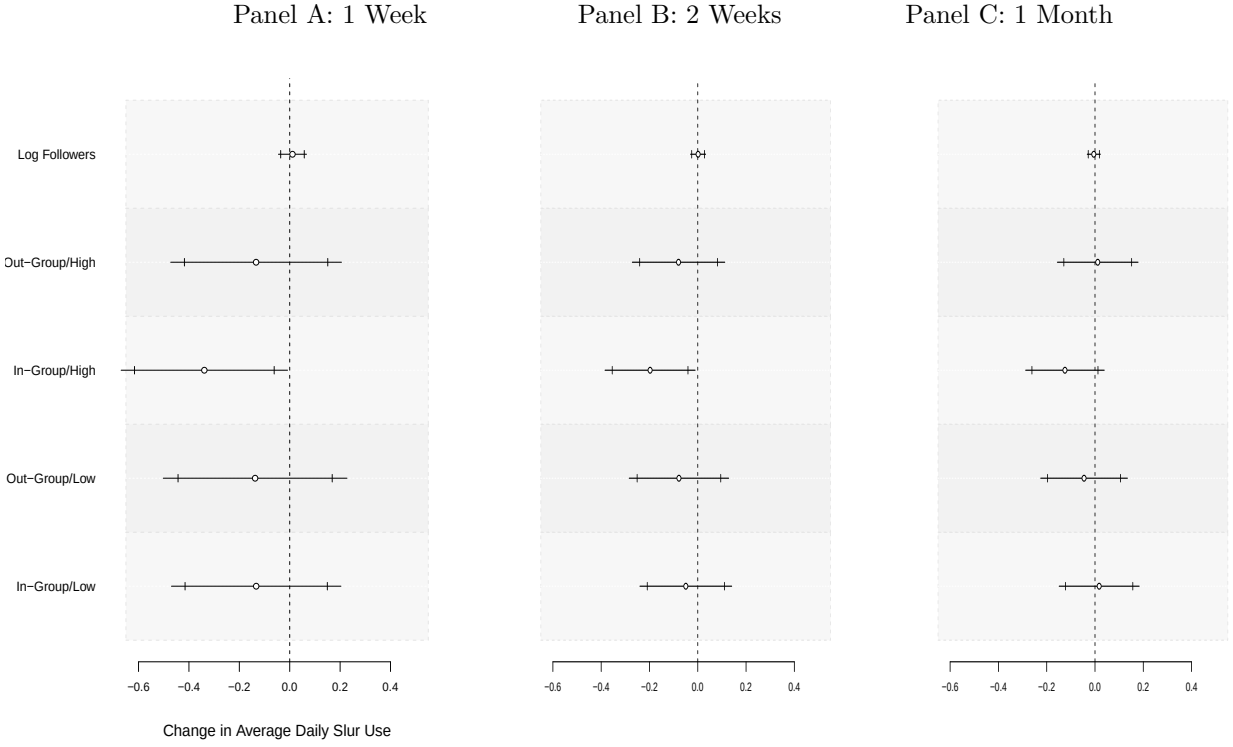
Each panel represents the results of a separate OLS regression in which the outcome variable is the absolute number of times a subjects tweeted the word "n****r" per day in the specified time period. For example, the coefficient associated with the In-Group/High Followers treatment in Panel A shows these subjects reduced their average daily usage of this slur by .26 more than subjects in the control in the week after treatment. Each regression also controls for the subject's absolute daily use of this slur in the two months prior to the treatment. The vertical tick marks represent 90% confidence intervals and the full lines represent 95% confidence intervals.

The only treatment that significantly decreased the rate of racist language use was the In-group/High Follower treatment. This is precisely what H_1 predicted to have the largest effect. There is a reduction in racist language use among the other three treatment conditions, but it is not significant at $p < .10$, and it is of smaller magnitude than the reduction in the In-group/High Followers condition. This was contrary to my expectations in H_1: I predicted that both the Out-group/High Followers and In-group/Low Followers conditions would have a larger effect than the Out-group/Low Followers condition.

Comparing across the panels of Figure 1.4 shows the decay in the effect of the In-group/High Follower over time. Although the effect remains statistically significant, the coefficient decreases steadily. In Panel A, the point estimate of -.26 indicates that the daily rate of the use of the word "n****r" decreased by .26 more among subjects in the In-group/High Follower Treatment condition than among subjects in the control condition. This average treatment effect decreased in magnitude to -.16 in Panel B and -.11 in Panel C. Treatment effects were not significantly different from zero after two months, so these results are not shown.

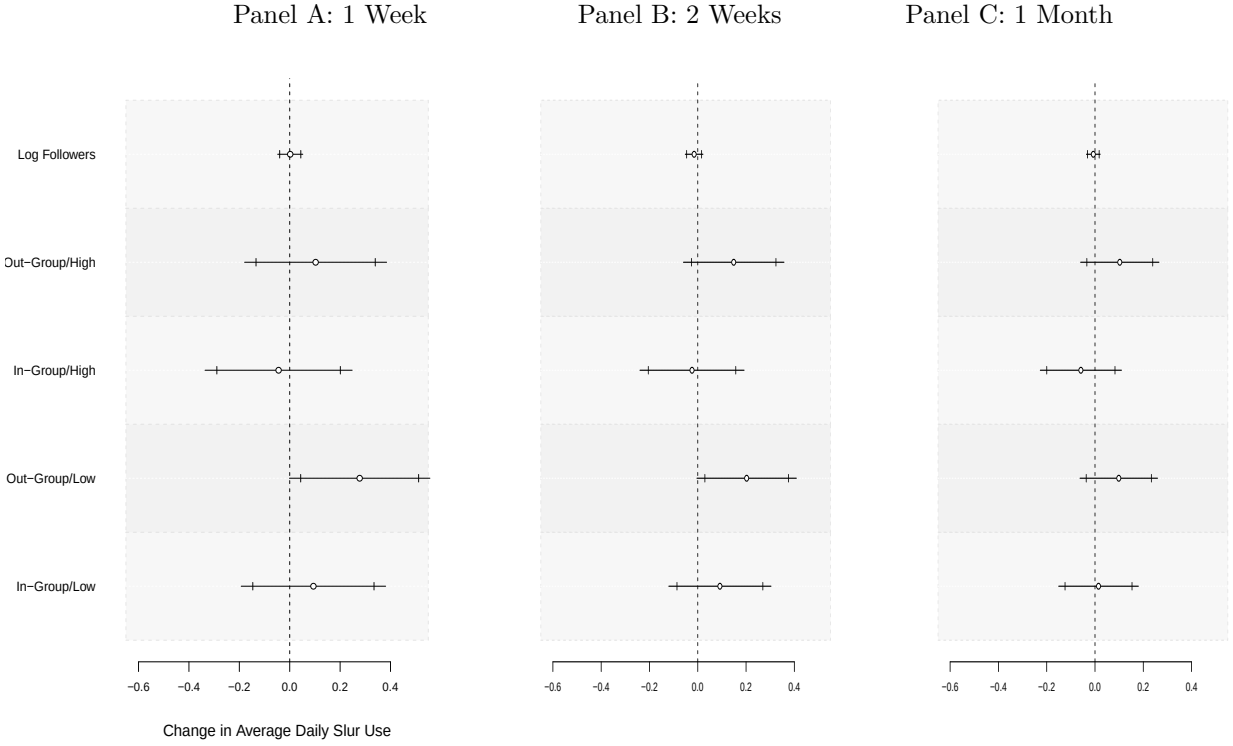
In order to test H_2, I divide the sample into two subgroups: those with Anonymity Scores equal to two, indicating that they shared no identifying information, and those with Anonymity Scores of either zero or one, indicating that they shared their real name, a real picture of themselves, or both. The anonymous sub-group had 159 subjects, and the non-anonymous subgroup had 84. Only 26 subjects had an Anonymity Score of zero, so I cannot divide this group further. My prediction in H_2 was that the reduction in harassment would be greater among the non-anonymous subgroup.

Figure 1.5: Reduction in Racist Slurs, Anonymous Subjects (*N=159*)



Each panel represents the results of a separate OLS regression in which the outcome variable is the absolute number of times a subjects tweeted the word "n****r" per day in the specified time period. For example, the coefficient associated with the In-Group/High Followers treatment in Panel A shows these subjects reduced their average daily usage of this slur by .34 more than subjects in the control in the week after treatment. Each regression also controls for the subject's absolute daily use of this slur in the two months prior to the treatment. The vertical tick marks represent 90% confidence intervals and the full lines represent 95% confidence intervals.

Figure 1.6: Reduction in Racist Slurs, Non-Anonymous Subjects (*N=84*)



Each panel represents the results of a separate OLS regression in which the outcome variable is the absolute number of times a subjects tweeted the word "n****r" per day in the specified time period. For example, the coefficient associated with the Out-Group/Low Follower treatment in Panel A shows these subjects increased their average daily usage of this slur by .28 more than subjects in the control in the week after treatment. Each regression also controls for the subject's absolute daily use of this slur in the two months prior to the treatment. The vertical tick marks represent 90% confidence intervals and the full lines represent 95% confidence intervals.

Figures 1.5 (anonymous) and 1.6 (non-anonymous) display the results. Figure 1.5 roughly mirrors the findings on the entire sample from Figure 1.4: there was a significant reduction among the subjects only in the In-group/High Followers condition, although in this case the effect was no longer significant after one month. However, the results from Figure 1.6 are starkly different. Not only was there no reduction in any treatment condition, there was actually a significant *increase* in racist language use among subjects in the Out-Group/Low Followers condition. This was the condition that H_1 predicted would experience the smallest reduction in racist language use, but the fact that this treatment caused an *increase* was surprising.

These results not only fail to support H_2, they provide evidence for the opposite conclusion: there was only a decrease in harassment among subjects with the highest Anonymity Score, and the *direction* of the effect changed (at least for one treatment arm) among subjects with non-maximal Anonymity Scores.

1.5 Discussion

The primary prediction expressed in H_1, that the In-group/High Follower treatment would cause the largest reduction in racist language use, was borne out. This effect was larger than either the In-group/Low Follower or Out-group/High Follower treatments, although these latter two reductions were not significant as expected. Overall, this is evidence of a *multiplicative* effect of the two treatments, as neither had an effect in isolation.

I found evidence for both Social Identity Theory in terms of in-group norm promotion and the theory that influential community members drive changes in normative group behavior. The sanctioning treatment caused subjects to update their beliefs about norms of online behavior, but only when the sanctioner was *both* a member of the in-group and perceived to be influential.

Encouragingly, these effects persisted for the first month under study, although not for two

months. Also, the p-value of the effect in the two week time period was actually smaller than for the one week and one month time periods. This non-monotonicity was surprising, relative to my expectation of a steady decay. My post-hoc explanation is that the smaller-than-expected effect sizes in the one week time period were caused by some subjects responding directly to the treatment by harassing the bot that tweeted at them and actively rebelling against the attempt to persuade them to change their behavior.

This phenomenon is called "reactance," and it has been shown to occur in a variety of political contexts. In a study of efforts to correct misperceptions, for example, Nyhan and Reifler (2010) find that, when confronted with evidence that a view they hold is false, some people actually become firmer in their false belief. More closely related to the current context, a study by Harrison and Michelson (2012) about eliciting donations to an LGTBQ organization finds that callers who self-identify as LGTBQ in an effort to personalize the issue are less effective than those who do not, and they believe that this is caused by reactance to the pressure implied by this personalization.

An example of reactance in my experiment is the subject who tweeted at my (black) bot twice: "@[bot] I DONT GIVE A FUCK N****R STFUFUCK YOU AND YOUR MOTHER" and "LMFAO N****R LOVERS NEEDA CHILL". For a subset of the subjects, reactance to the treatment actually caused a short-term increase in the use of racist language. Only around 30% of the subjects responded to the treatment, and this rate did not vary across the treatment arms.[17] Overall, this phenomenon was overwhelmed by the overall decrease in the longer time periods. Future studies should employ a larger sample size to better differentiate between these short- and long-term effects of social sanctioning.

The effect of Anonymity was found to be contrary to my prediction in H_2. My expectation was that the treatment effect would be smaller for more anonymous subjects, as suggest by the findings in Omernick and Sood (2013) and Hosseinmardi et al. (2014). However, the treatment effects turned out to be smaller among *less* anonymous subjects, and the

[17]These responses also did not vary in terms of vitriol between the treatment arms. In fact, even the number of subjects that responded to call my bot a "n****r" did not vary significantly between the white and black bots.

treatment caused an *increase* in harassment for the non-anonymous subjects in the Outgroup/Low Followers condition. This is consistent with the expectations of SIDE theory, and with the findings in Postmes et al. (2001).

Still, "Anonymity" in the current context does not map exactly onto these previous findings, and I urge caution in generalizing the results of this study. Specifically, these subjects *selected their own level of anonymity*, according to some process that is not well understood. The heterogeneous treatment effects may not represent the effects of anonymity *per se*, but of some other unobserved characteristic of the subjects. Future research on why people choose to remain anonymous on mixed-anonymity platforms like Twitter can help solve this puzzle.

1.6 Conclusion

Online communities represent an important development in empowering people to express themselves and communicate with the world without being limited by their physical location or social status. However, this freedom also enables some individuals to behave badly, unconstrained by social norms and uninhibited by biological feedback mechanisms restricting antisocial behavior. One manifestation of this is the harassment of members of disadvantaged groups, aiming to silence and weaken the victims of this harassment and to solidify in-group membership. In the context of the US, this often takes the form of white men harassing women and racial minorities.

To address this problem, online network administrators or government entities can explicitly ban harassing individuals or restrict certain language use. These efforts can backfire, though, and cause people to use even more racist or misogynist slurs to better differentiate themselves and their group from the "Political Correctness" they associate with censorship. Approaches that operate through promoting positive social norms, like the one employed in this paper, may offer a better way to develop online communities that are less toxic.

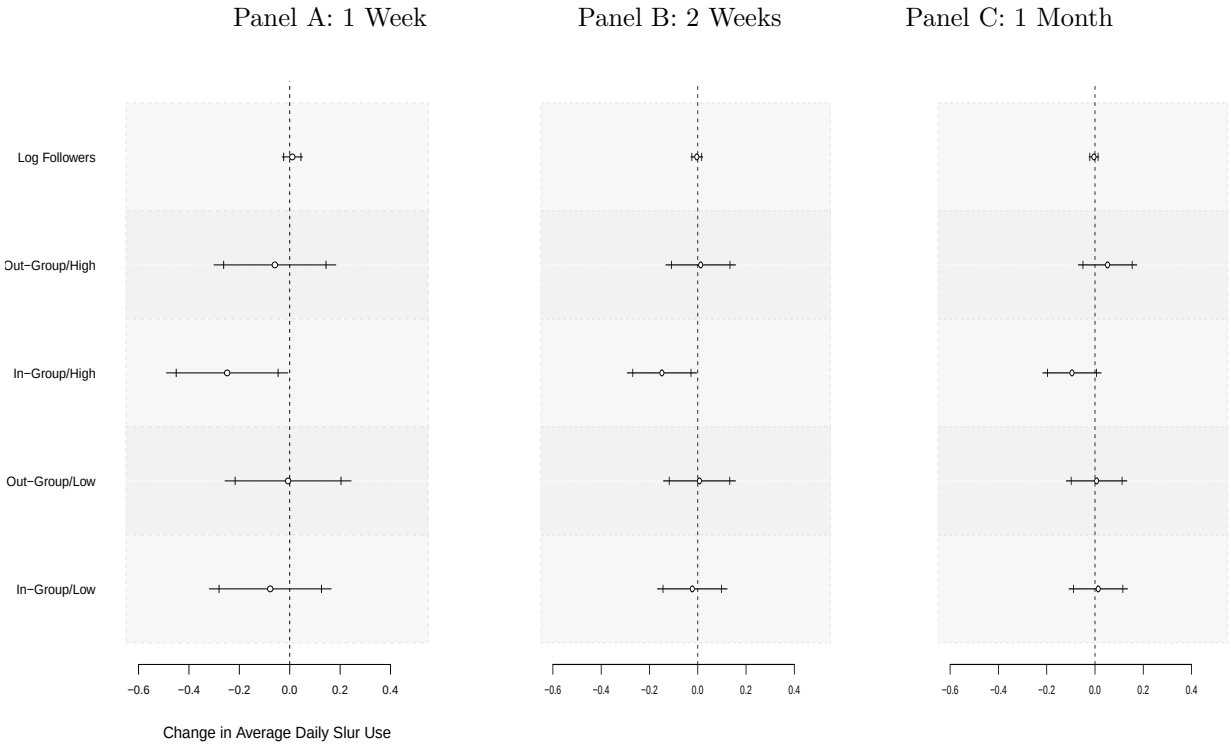
The experiment performed in this paper tests another approach to reduce the incidence of racist online harassment. By explicitly priming the subjects' membership in offline communities and updating their beliefs about the norms of online behavior, the treatment caused a significant reduction in the use of racist slurs. However, this effect was only observed among the subsample of subjects who had anonymous profiles. Among subjects who disclosed personal information, there was no significant reduction in the use of racist slurs, and there was actually an increase in the use of racist slurs among one treatment condition.

Although prejudice reduction has been widely researched, previous studies have been limited by a combination of convenience samples of undergraduate students, self-reported outcome variables, and a short measurement period that cannot measure effect persistence. Following Paluck and Green (2009)'s call for more randomized field experiments in prejudice reduction, this paper represents an improvement in all three of these dimensions: the subjects were drawn from the general population and selected because they engaged in public harassment, the outcome variable was behavioral and objective, and the measurement period was continuous and two months long.

This method, of performing experiments on subjects on social media using accounts the experimenter controls, can be applied to many contexts in which the outcome of interest is online speech. An important extension to this study would be a manipulation to reduce misogynist online harassment, which continues to be a large problem for women on social media. More broadly, it could be used to experimentally determine the best method to dissuade people on social media from communicating false and potentially dangerous information about, for example, vaccinations. However, the findings from this study do not trivially generalize to offline communication or behavior.

Although this study's demonstration of a method to reduce the expression of prejudice online is valuable in and of itself, the question remains as to whether this effect changes underlying prejudiced attitudes or behavior in the physical world. Ideally, future contributions in this area of study should aim to measure all three out of these outcomes.

Figure 1.7: Reduction in Racist Slurs, Conservative Assumption



Each panel represents the results of a separate OLS regression in which the outcome variable is the absolute number of times a subjects tweeted the word "n****r" per day in the specified time period. For example, the coefficient associated with the In-group/High Followers treatment in Panel A shows these subjects reduced their average daily usage of this slur by .25 more than subjects in the control in the week after treatment. Each regression also controls for the subject's absolute daily use of this slur in the two months prior to the treatment. The vertical tick marks represent 90% confidence intervals and the full lines represent 95% confidence intervals.

Appendix for Chapter 1

A.1: Conservative Assumption for Main Results For the subjects who produced too few post-treatment tweets to calculate an rate of racist language use, I assumed that their post-treatment rate of racist language use was zero. This assumption makes sense substantively, because these people were no longer tweeting (and thus no longer engaging in racist harassment). However, a more conservative assumption would be to assume that there was no change in their behavior, and to assign them a post-treatment rate equal to the their pre-treatment rate. This does not substantively change the results, although the magnitude of the effect sizes becomes slightly smaller.

Chapter 2

Don't @ Me: Experimentally Reducing Partisan Incivility on Twitter

2.1 Introduction

In October of 2016, President Obama claimed (and Democratic Presidential nominee Hillary Clinton tweeted) that "civility is on the ballot." Concern over political civility was widespread during the 2016 US Presidential election, and many felt that the internet and social media (which Republican Presidential nominee Donald Trump employed enthusiastically) were to blame.

Concern over incivility in contemporary political discourse can be traced back at least to the rise of cable news and its personalistic, outraged style (Berry and Sobieraj, 2013; Mutz, 2015). Indeed, concern about civil discourse may accompany any technological advance that lowers the cost of information production and distribution; the invention of the printing press led to elite concern about civil discourse during the time of the Reformation (Bejan, 2017).

Modern technological changes are taking place in the context of increased partisan animosity. Often called "affective polarization," this animosity reflects a growing distrust and lack of respect among Democrats and Republicans (Iyengar, Sood and Lelkes, 2012). This phenomenon is directly related to civility, which Mutz (2015) says is "a means of demonstrating mutual respect" (2015, p7). Incivility is more than impoliteness: it is indicative of a disregard for the act of deliberation. Internet technologies may or may not be driving affective polarization, but they do at a minimum allow for the lack of mutual respect to manifest itself in incivil online discourse.

Online communication lacks the evolved social and emotional feedback mechanisms that make it difficult to be incivil in a real-world setting, and it affords physical distance and (sometimes) anonymity, decreasing the effectiveness of social sanctioning (Frijda, 1988). These technological affordances, in a context replete with bad actors intent on sowing discord for fun (Phillips, 2015) or geopolitical advantage (Chen, 2015), have degraded norms of civil discourse online.

The implications of these changing norms are serious. Early enthusiasm about the capacity of the internet to democratize political discourse may have been premature (Hindman, 2008), but the affordances of today's ubiquitous, easy-to-use and social internet have caught up with the hype: in 2008, only 25% of US adults were on a social network, but during the 2016 US Presidential Campaign, that number was 68% (Greenwood, Perrin and Duggan, 2016).

Ideally, the internet could enable broad, direct, deliberative democracy, with all of the desirable normative qualities this entails. Technological advances may continue to expand the breadth and depth of online communication, but if incivil discourse remains the norm, deliberative democracy will remain out of reach.

Deliberative democracy entails more than mere communication. It only works to the extent that participants sincerely weigh the merits of the arguments being deliberated and that this consideration is not contingent on the identity of the person making the argu-

ment (Fishkin, 2011). In a context of high affective polarization and norms of partisan incivility, this rhetorical charity does not obtain. Survey evidence suggests that internet users do not feel that their interactions are deliberative–"64% say their online encounters with people on the opposite side of the political spectrum leave them feeling as if they have even less in common than they thought" (Duggan and Smith, 2016).

I conducted an experiment that evaluates different strategies for promoting civil political discourse during the 2016 US Presidential election. Using the method developed by Munger (2017*c*), I used Twitter accounts that I controlled to sanction users engaged in incivil discussions. In contrast to lab experiments conducted on a convenience sample in a short time frame, this approach allowed me to measure the effectiveness of sanctioning on a sample of frequently incivil partisans in a realistic setting and in a continuous and unbounded time frame.

Users were sampled by searching for tweets that mentioned either @realDonaldTrump or @HillaryClinton but which were directed at another, non-elite user. Using an algorithm developed to identify aggression in comments on a Wikipedia editors' discussion forum (Wulczyn, Thain and Dixon, 2017), I selected the tweets most likely to be incivil. I then manually inspected the interaction to ensure that it was a true instance of a non-elite[1] being incivil to another non-elite of the opposing partisan persuasion. I then randomly assigned the subject to a treatment arm–subject to the balance constraint that each treatment pool have the same distribution of subject Anonymity Scores–and used "bots" to send them a message.

By manipulating the partisan identity of my "bots,"[2] I test the differential effects of sanctioning on Republicans and Democrats. By varying the language I tweeted at subjects, I test hypotheses about the relative effectiveness of two kinds of moral suasion and include a non-moral message that simply reminds subjects that what they are tweeting is public.[3] All

[1]I define as an "elite" anyone who was "Verified" on Twitter–they had a blue check mark next to their name which means that Twitter has verified that they are who they say there, a status which Twitter only bestows on users they consider public figures–or anyone who identified themselves as a journalist or political operative in their profile.

[2]These are not "bots" in the sense that they behave autonomously; I did all of the tweeting manually. I refer to them as bots throughout the paper for lack of a better term.

[3]The research design, dependent variable measurement, and main hypothesis were pre-registered at EGAP.org prior to any research activities.

treatment effects are calculated as an individual-level change in tweeting behavior relative to a true control group that did not receive any treatment; this approach differences out any effects of real-world events on subjects' behavior.

I found evidence of significant changes in subjects' behavior, but the effect heterogeneity took an unexpected form. There was no difference between the effectiveness of the two kinds of moral suasion, but there was a significant difference between Democrat and Republican subjects: Democrats significantly reduced their rate of incivility in response to either moral treatment, but Republicans did not change their behavior for either, in the primary time period of the first week post-treatment.

In the first *day* post treatment, however, the reverse was the case: one of two moral treatment conditions caused a significant reduction in incivility among Republicans, but neither had an effect on Democrats. This contrast emphasizes the importance of research designs which are capable of measuring effect persistence.

Subject anonymity significantly moderated treatment effects, in the expected direction: more anonymous subjects were less likely to respond to the treatment. This trend was only significant in the one-day time frame.

I also theorized that both moral treatments would be more effective at reducing incivility than a non-moral message that reminded users that what they were saying was public. This trend was not statistically significant overall, and among Republicans, the moral and non-moral messages had precisely the same effect. Among Democrats, however, the expected difference was highly significant.

These findings demonstrate that various different forms of moral suasion can be effective in promoting a more civil political discourse on Twitter, above and beyond the effect of merely calling attention to the subjects' behavior. This moral suasion may only be effective on a subset of users; anonymous users (those more likely to be trolls) were unresponsive to moral suasion, and may even have been encouraged by being told that they were violating norms of political civility. Efforts to promote online civility should be sure to target the

right people and use the most appropriate rhetorical strategy to maximize their efficacy.

2.2 The Promise and Perils of Social Media

Perceptions of the impact of social media (and the internet more generally) on democratic politics have changed dramatically in the brief period of social media's existence. Initial optimism suggested that citizens would be better able to communicate with both their governments and with each other, unconstrained by geography and the power imbalances of the physical world (Papacharissi, 2002). Although conversations could get heated and impolite, the overall effect was to revitalize the public sphere of debate (Papacharissi, 2004). The campaign manager for Howard Dean, one of the first politicians in the US to fully embrace the power of the internet for politics, said that "the internet is the most democratizing innovation we've ever seen, more so even than the printing press" (Trippi (2004), quoted in Hindman (2008)).

Indeed, a wide variety of politicians began using social media to communicate with their constituents (Gulati and Williams, 2010). Individual politicians are better able to reach voters directly, rather than through the mediating institution of party control (Karlsen and Skogerbø, 2013). Although the process does not always work perfectly, there is evidence that politicians respond to the citizens who engage with them on social media, discussing topics that citizens bring to their attention (Barberá et al., 2014). Additionally, citizens do seem to learn about party platforms directly from communication by politicians on Twitter (Munger, Egan, Nagler, Ronen and Tucker, 2016).

On the non-elite side, the use of the internet to discuss non-political topics has enabled some cross-cutting ideological mass discussion (Wojcieszak and Mutz, 2009). This phenomenon first began with blogs. By 2006, 8 million US citizens claimed to share their thoughts through online blogs, and fully 57 million US citizens claimed to read them (Hindman (2008), p104). Hindman describes the prevailing mood at that time, when media

commentators were lauding the development of blogs as a brave new world for deliberative democracy: "The central claim about blogs is that they amplify the political voice of ordinary citizens." However, as he argues persuasively in *The Myth of Digital Democracy*, the infrastructure of the internet tends to lead to an even more skewed distribution of readership than does traditional media: "It may be easy to speak in cyberspace, but it remains difficult to be heard. (p142)"

When the competition to be heard is intense, competitors often resort to using outrageousness to garner attention. For example, when cable enabled new entrants to the television marketplace, these upstart media organizations were willing to blend news and entertainment in a way that traditional network broadcasters had resisted. In the words of Bill O'Reilly, host of the famously confrontational television program *The O'Reilly Factor*: "The best [cable news] host is the guy or gal who can get the most listeners extremely annoyed over and over and over again" (O'Reilly (2003), cited in Mutz (2015)). Norms of journalistic integrity established in the early 20th century rapidly eroded, resulting in less civil media and citizens who trusted and liked that media less (Berry and Sobieraj, 2013; Ladd, 2011).

This tendency was exploited during the 2016 Republican Primary and Presidential Election by Donald Trump. In the midst of the Primary, the *New York Times* estimated that Trump had benefited from nearly $2 billion worth of free media coverage on television; despite spending less than his opponents on ads, he ended up with far and away the most airtime (Confessore and Yourish, 2016). By treating his opponents derisively and incivilly, he captured the attention of the media, whose consumer-financed business model meant they profited by catering to the desire of the public to consume news coverage of this incivility. His use of Twitter as the medium for incivility meant that he could influence the news cycle for free, from anywhere, and at strategically useful times (Francia, 2017). As Lawrence and Boydstun (2017) put it, "Trump's entertaining, sensational, inflammatory words and actions make him the kind of phenomenon we just can't look away from." The resulting prevalence of Trump's incivil behavior—and the apparent success of the strategy in winning him the

Presidency—has eroded norms of civility among the political elite.

An analogous trend took place in citizen online engagement, but earlier, more rapidly, and to a greater extreme; in contrast to incivility in cable news, incivility in online discussions can be explicitly targeted at someone else. Early forums tended to be anonymous, and early internet users flocked to sites like 4chan and somethingawful to discuss whatever was on their mind. However, a subset of these people found that this anonymity empowered them to say incivil and outrageous things, and that they could easily upset other users. This behavior soon spread over the internet, as "trolls" mocked memorial pages on Facebook and posted vivid images of gore and hardcore pornography so that other users might suffer serious emotional turmoil (Phillips, 2015).

This kind of behavior is enabled by Computer Mediated Communication (CMC). In the physical world, evolved social and emotional feedback mechanisms make it emotionally difficult to look a stranger in the eye and say something incivil (Frijda, 1988), but these mechanisms are lacking in CMC, as are physical proximity and identifiability. The overwhelming majority of social media users report this experience: 84% of social media users say that people say things when discussing politics online that they would never say in person (Duggan and Smith, 2016). CMC makes it difficult to enforce social norms, and while this does tend to encourage more communication and creativity, it also allows even a small number of ill-intentioned actors to impose significant emotional costs on other users (Bordia, 1997; Kiesler, Siegel and McGuire, 1984; Walther, 1996).

The competition for attention and the difficulty of punishment in anonymous contexts meant a race-to-the-bottom in terms of online speech norms. Today, the internet is widely regarded as rife with offensive and even harassing speech designed to mock sincere expression–trolling culture is dominant online (Buckels, Trapnell and Paulhus, 2014; Milner, 2013). The extent to which trolling culture obtains, though, depends on the specific technical affordances of different online platforms. The most important feature, in this respect, is the extent to which platforms allow their users to be anonymous. Studies have consistently found

that more anonymous platforms experience more harassment (Hosseinmardi et al., 2014; Omernick and Sood, 2013).

Facebook, for example, has invested heavily in linking their users' accounts with their real identities. Twitter, on the other hand, allows all manner of parody, comedy and anonymous accounts. Twitter has consistently defined itself as in favor of free speech, and while this has made it the preferred platform for revolutionaries in both Western countries and authoritarian regimes around the world (Barberá, Wang, Bonneau, Jost, Nagler, Tucker and González-Bailón, 2015; Earl et al., 2013), it has also become notorious for failing to curtail harassment. In the candid words of Twitter's CEO Dick Costelo in an internal memo in 2015, "We suck at dealing with abuse and trolls on the platform and we've sucked at it for years." There is evidence of the direct implications of the affordances of a given platform and its capacity for deliberation: Halpern and Gibbs (2013) show that less anonymous platforms like Facebook are hosts to more deliberative interactions, whereas more anonymous platforms see a higher incidence of impolite comments. Interventions on Facebook have been shown to be effective in decreasing the incivility of online commenters, especially when the intervention is performed by an identifiable person rather than a faceless institution (Stroud et al., 2014).

2.3 Partisan Incivility, Affect Polarization and Deliberation

The development of social media as both a platform for political communication and a locus for incivility took place at the same time as a sharp growth in animosity between Democratic and Republican partisans. Scholars have described this trend as "affective polarization"– partisans dislike each other (Iyengar, Sood and Lelkes, 2012) and tend to trust co-partisans and distrust out-partisans more (Iyengar and Westwood, 2015). This phenomenon has even extended to the marriage market, as preferences for a partner with similar partisan characteristics today are roughly half as large as preferences for a partner of the same race

(Huber and Malhotra, 2017).

Although the uptick in partisan polarization began well before the mass adoption of social media, there exists a plausible connection between the two. Some scholars claim that social media use exposes people to a wider range of views and thus decreases issue polarization (Barberá, 2014), but others argue that social media inflames partisan emotions and increases affective polarization (Settle, Forthcoming). The large-scale, contemporaneous development of social media and affective polarization makes causal claims difficult to establish; an exception is Lelkes, Sood and Iyengar (2015), who use the quasi-random rollout of broadband internet as an instrument for the use of social media and find that it significantly increased affective polarization.

Regardless of causality, it is clear that incivil political arguments take place on social media. Sometimes the incivility is directed at politicians themselves, and while we might expect that having a thick skin is necessary to survive in that business, Theocharis et al. (2015) show that this can decrease politician engagement with their constituents on Twitter. Perhaps more importantly for the mass public, this behavior means that citizens who wish to engage with politicians or each other in response to a politicians' tweet are necessarily exposed to incivil messages. The presence of incivility thus has a *compositional* effect on online political discourse: people with a low tolerance for incivil discourse are likely to disengage from politics if the norm is of incivility. 39% of social media users report having changed their settings or removed a user from their feed because someone posted something offensive, posted too often or posted something abusive (Duggan and Smith, 2016). There also appears to be a *direct* effect of incivility on an individual's discursive style: Cheng et al. (2017) find that discussants who join an online forum and see that an incivil discussion is taking place are more likely to be incivil themselves. These two effects have allowed the norm of incivility in online discussions set by a small group of committed trolls to spread to the broader online community.

Incivility comes far more naturally if you believe your interlocutor deserves it; incivility is

entailed by increasing affective polarization. I follow Mutz (2015): "Following the rules of civility/politeness is...a means of demonstrating mutual respect" (2015, p7). If mutual respect between partisans is decreasing, it should be no surprise that civility in their conversations is decreasing as well.

The implications for deliberative democracy are serious; Fishkin's model claims that deliberative democracy works to the extent that participants sincerely weigh the merits of the arguments being deliberated and that this consideration is not contingent on the identity of the person making the argument (Fishkin, 2011). This does not at all describe the dominant mode of political discourse online: rather than leading to an exchange of information and arguments that can potentially lead to consensus, a name-calling match between partisans online may actually cause both parties to think less of their opponents and their arguments, driving the parties even further from consensus. Experimental evidence supports this theory: exposure to uncivil online discussion increases perceptions of polarization and makes subjects less likely to experience a deliberative discussion (Hwang, Kim and Huh, 2014).

I do not mean to imply that the ideal level of incivil speech, or that consensus is the only normatively desirable outcome of speech. Incivility serves an important function, allowing people to call out intolerable behavior and to express themselves. Irrespective of its desirability in the abstract, Sanders (1997) argues that deliberation in the extant, unequal American society inherently privilege the voices of one class (and race, and gender) of people over others. An alternative model of democratic speech is one that values "giving testimony," the expression of voices which would otherwise be excluded.

"Giving testimony" has become radically easier with the advent of social media, to the benefit of the public. Never before in American history has there been as much diversity of discussion as on Twitter in 2016. This diversity, though, is a step away from consensus, a troubling fact often elided by early deliberative democratic theorists when they talk about humans in the abstract (Sanders, 1997). In fact, from the perspective of political discourse, the widespread adoption of social media over the past ten years is likely the most radical step

towards enabling giving testimony in the history of humanity. For a summary of scholarship on the role of diversity in deliberation, see Bächtiger et al. (2010).

The normative case for reducing online partisan incivility, then, is that there was simply too much of it during the 2016 US Presiential ELection: Computer-Mediated Communication has decreased the cost of being incivil online, partisan affective polarization has increased the demand for incivility in partisan discourse, and the campaign of Donald Trump eroded norms of elite discourse. Although deliberation is not the only valuable form of political speech, it is a central function of political speech in a democracy, and it was difficult to realize on Twitter during the 2016 US Presidential Election.

2.4 Experimentally Reducing Political Incivility

I conducted an experiment to sanction users who were sending incivil messages to out-partisans and measured the change in their behavior.

The first step in performing this experiment was finding conversations that were incivil, between out-partisans, *and* about politics. I thus used streamR to scrape the streaming Twitter API for tweets mentioning either "@realDonaldTrump" or "@HillaryClinton"—the Twitter accounts of the two major party candidates in the 2016 US Presidential election. I then dropped any tweets that were not directed at another user who was *not* either Trump or Clinton.

In this way, I found a sample of tweets from non-elites that were concerned with the "issues" most likely to inspire political incivility in October 2016: Trump and Clinton. In order to filter through the hundreds of thousands of tweets every hour that fit these criteria, I used a machine learning classifier designed to detect aggression. Wulczyn, Thain and Dixon (2017) trained and evaluated a neural network on millions of comments on Wikipedia "talk pages" (the behind-the-scenes part of Wikipedia where editors discuss potential changes) in a format that is reasonably similar in structure and length to Tweets. The algorithm

Figure 2.1: Finding Non-Elite Incivility

performed very well on Wikipedia comments, with an AUC score of 96.6—higher than taking the majority vote of three human annotators.

I used the model to assign an "aggression score" to each tweet I had scraped, then manually evaluated the top 10% most aggressive tweets per batch.[4] From these prospective subjects, I selected the ones who were directing incivil language at a member of the opposite political persuasion. Many of the potential subjects I found this way were tweeting at elites–either people verified on Twitter, journalists or campaign operatives–and I excluded them. I also found many people agreeing (though often in incivil ways) with an in-partisan about how terrible the out-party is, and excluded them as well. When performing a manual inspection of the potential subject's profile, I excluded users who appeared to be minors or who were not tweeting in English. I also checked to ensure that the subject's profile was at least two months old; Twitter does ban some user accounts for harassment or other violations of their Terms of Service, so a very new account is likely to have been started by someone who had previously been banned. A new user is also likely to have too short a tweeting history for me to establish a reasonable baseline for their past behavior.[5]

[4]This process was time-consuming, and there were a finite number of tweets satisfying my criteria being tweeted at a given time, so I iterated this scrape-validate-treat procedure several times.

[5]This process was pseudo-algorithmic: these are the criteria I set out for myself in selecting subjects. I was the sole coder in making these decisions, and it is possible that some misclassification may have occurred and I may have included in the subject pool a Twitter user who did not meet my explicit criteria: they might not have been a frequently incivil partisan who had just tweeted an incivil message about politics at someone of the opposite persuasion. There is no "ground truth" here, so it is impossible to know how often this happened, but any such errors in subject classification do not pose a problem for causal inference due to the random assignment of treatment.

For a visual overview of this selection process, see Figure 2.2. In this way, I found incivil tweets from a non-elite to another non-elite with whom they disagreed politically. For an example, see Figure 2.1. @realDonaldTrump tweeted something, then Parker tweeted "you already lost" at Trump.[6] Ty then responded to Parker (but because of how Twitter works, Ty's tweet also "mentions" @realDonaldTrump) with an incivil comment. Ty is the subject I included in the experiment, and because he was being incivil to someone criticizing Trump, I coded Ty as a Trump supporter.

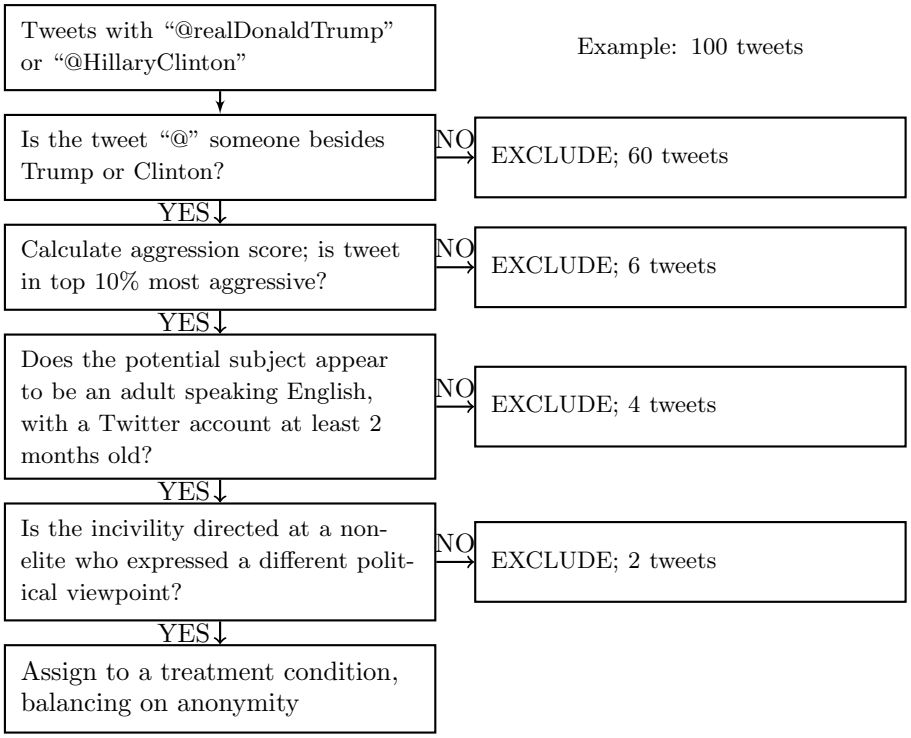
The sample in this experiment is thus not a "representative sample" of the general population, or even of Twitter users. The group of people encountered by following the steps listed in Figure 2.2 are, however, precisely the type of Twitter users who might be able to deliberate with and learn from their political opponents, if they were to do so in a civil fashion.

Based on findings in Munger (2017*c*), and on the theoretical expectation that anonymity is an essential part of what enables incivility online, I also recorded each subject's Anonymity Score during the subject discovery process. The Anonymity Score ranged from 0 (least anonymous, full name and picture) to 2 (most anonymous, no identifying information). Ty, from Figure 2.1, was coded as a 1–he chose to display what could plausibly be his full name. He also provided some personal information in his "bio" field, to the left of where he claims to be an "All around nice guy!", which I censor for privacy reasons.

My aim was to convince subjects that they were being sanctioned by a real person, so I made my bots look as real as possible. After I tweeted at a subject, they received a "notification" from Twitter. Non-elites are unlikely to get more than a few notifications per day, so they almost certainly saw the message I sent them. It is uncommon to be tweeted at by a stranger, but not extremely so, and especially not among a subject pool who are tweeting incivil things at out-partisans. As a result, they were likely to click on my bots' profile; if they did, they would see something very like Figure 2.3.

[6] I censor the usernames of the subjects to preserve their anonymity. In principle, the exact text of a tweet should be enough to find a user, but the phrases used in this exchange are quite common.

Figure 2.2: Sample Selection Process



This flowchart depicts the decision process by which potential subjects were discovered, vetted and ultimately included or excluded. The right-hand column gives an example of the number of potential subjects excluded at each step in the process.

Figure 2.3: (a) Example Bot–Clinton Condition

(b) Example Bot–Republican Condition

Neil, in panel (a), was a bot who appeared to be pro-Clinton. I created four bots; the other three were pro-Democrats, pro-Trump, and pro-Republicans (see Todd, in panel (b)). To manipulate these identities, I changed the large banner in the middle of the profile, the small logo in the bottom right of the bots' profile pictures, and the "bio" field below their username (eg "Hillary 2016!"; "Republicans 2016!"). The four bots were otherwise identical. All of the bots appeared to be white men, keeping the race/gender aspect of the treatment constant. I used identical cartoon avatars to avoid anything about the users' appearance priming the subjects; it is not uncommon for Twitter users to have cartoon avatars, so this was unlikely to raise suspicions.

I took other steps in order to maximize verisimilitude. Most importantly, I ensured that all of the bots had a reasonably high number of followers. Munger (2017*c*) varied the number of followers that sanctioning bots had, and found that bots with few followers had very little effect. Based on this finding, I purchased 500 followers for each of my four bots, although each bot actually got 900 followers. The number did not vary significantly among the four.

I created each bot in January 2015, giving the impression that they were long-time users. When creating the accounts, I followed Twitter's recommendation to follow 40 pre-selected accounts, mostly celebrities and news services. To further increase the perception that the bot was a real person, I tweeted dozens of innocuous observations (eg "I'm thinking of pasta for lunch.....YUM") and retweeted random (non-political) stories from the accounts the bots followed.[7]

There were two subject pools: people who were incivil to people critical of Trump ("Republicans") and people who were incivil to people critical of Clinton ("Democrats"). Within each of these pools, each subject was randomly assigned one of three messages ("Feelings", "Rules", or "Public") sent by one of two bots (pro-candidate or pro-party). There were initially 118 subjects in the "Republicans" pool, 104 subjects in the "Democrats" pool, and

[7]Bizarrely, the followers I bought sometimes "liked" and even occasionally retweeted these observations. These followers were reasonably realistic, and it is unlikely that anyone who looked at the bots' followers would realize they had been purchased.

another 108 in the control group, to whom I sent no tweets.[8]

The primary outcome of interest was how subjects responded to being sanctioned, both in terms of their direct response to the sanctioning tweet and in how they changed their behavior after having been sanctioned. I only used bots that appeared to be on the same "side" as subjects to send the sanctioning message; I was concerned that cross-ideological sanctioning might cause subjects to react angrily and send even more incivil messages. I had no theoretical expectation as to whether right-leaning or left-leaning subjects would respond more to being sanctioned.

The primary variation in the treatments was in the language of the message sent to the subjects. The aim was to convince subjects that their behavior is wrong–or at a minimum, to convince them to change their behavior. One approach, the one employed in Munger (2017*c*), is *in-group social norm promotion*: to cause subjects to update their beliefs about correct normative behavior for someone sharing their social identity. Munger found that sanctioning from bots that shared a social identity with the subject was more effective in changing their behavior than bots with a different social identity. To build on this finding, I held in-group social identity (in this case, partisanship) constant in the current study.

By varying the language of the in-group sanctioning, I tested the possibility of moral suasion. I based my approach on the moral intuitionist model proposed by Haidt (2001), which argues that moral emotion is antecedent to moral reasoning. People make moral judgments based on deep-seated intuitions and then justify those judgments with ad hoc reasoning. As a result, moral appeals should be targeted to these fundamental intuitions, rather than to the putatively logical justifications for specific judgments.

Extending the theory, (Haidt, 2012) argues that a necessary component for moral suasion is convincing your interlocutor that you are sympathetic and understanding. If the two of you share the same fundamental moral intuitions, you can reasonably discuss specific implications of those foundations, but if not, attempts to change their mind are likely to be

[8]In the analysis below, I include 310 subjects out of this original pool of 330. I discuss the attrition process in Appendix A.

interpreted as attacks on their worldview and to be met with resistance. To this end, all of my messages begin by identifying my bot and the subject as members of the same party (Democrat/Republican).

Haidt also finds that the morality of liberals and conservatives rests on different foundations. He finds six dimensions of morality that seem to operate in cultures around the world: Care, Fairness, Liberty, Loyalty, Authority, and Sanctity. These foundations capture a large portion of what constitutes "morality" across cultures, though they are not exhaustive. He argues that people in non-Western societies are similar to conservatives in the West in that both groups appear to place significant weight on all six of these moral foundations. Westerners on the left of the political spectrum, however, appear to put far more emphasis on just two: Care and Fairness.

As a result, liberals and conservatives speak past each other on some moral issues. For example, liberals sometimes have difficulty understanding why conservatives are so upset about flag burning. Burning a flag does nothing to cause harm (the primary question underlying the Care foundation), nor is it unfair, so liberals tend not to see it in moral terms. Conservatives, though, feel that it is disloyal and disrespectful to authority, and that flag burning is thus immoral.

Haidt's theory thus predicts that attempts at moral suasion will be more effective if they appeal to the favored moral foundations of the interlocutor. To that end, I designed two different treatments. The first was designed to appeal to the Care foundation, and thus to have some effect on Republicans but a much larger effect on Democrats[9]:

```
@[subject] You shouldn't use language like that. [Republicans/Democrats]
need to remember that our opponents are real people, with real feelings.
```

The other treatment appealed to the Authority foundation.[10] My expectation was that

[9]In *The Righteous Mind*, Haidt argues that Democrats' morality is built on Care, but specifically on care for certain victim groups who have traditionally been marginalized in US society. This treatment would thus be less effective if Democrat subjects perceive their Republican interlocutors to not be deserving of care.

[10]The specific language used in this treatment does not comport exactly with Haidt's conception of Authority; it lacks

it should have an effect on Republicans but not on Democrats:

```
@[subject] You shouldn't use language like that. [Republicans/Democrats]
need to behave according to the proper rules of political civility.
```

In addition to these moral foundations treatments, I included a non-moral "public" treatment. My intention was to use a message that would serve to remind subjects that their incivil tweets were public, and my hypothesis was that this treatment would decrease the subjects' use of incivility, but that the effect would be smaller than the moral treatments'. To that end, I designed a message that emphasized the subject's visibility:

```
@[subject] Remember that everything you post here is public. Everyone
can see that you tweeted this.
```

Hypothesis 3 *The reduction in incivility caused by the Care condition will be larger for Democrats than for Republicans. There should be a reduction in incivility caused by the Authority condition for Republicans, but not for Democrats. There should be a reduction in incivility cased by the Public condition, but it should be smaller than the other effects.*

Some subjects are more heavily invested in their online identities than are others. Twitter allows individuals to decide how much personal information to divulge, so while some users are completely anonymous, others include their full name, picture, and biography. There are likely to be large differences in how these different types of users engage with Twitter. Users who are more invested in their online identities are more likely to change their behavior in response to sanctioning, while anonymous users are unlikely to do so.[11]

Hypothesis 4 *The reduction in incivility caused by the treatments will positively covary with the subject's Anonymity Score.*

an explicit connection to hierarchy. A more theoretically consonant message would invoke some conception of, for example, American Traditions or the office of the President.

[11] Note that this hypothesis was not recorded in the Pre-Analysis Plan, but follows directly from the theory in Munger (2017*c*).

2.5 Results

The behavior targeted in this experiment is partisan incivility targeted at other Twitter user. To capture this behavior, I scraped each subject's Twitter history before and after the treatment and restricted the sample to the tweets that were "@-replies": tweets directed at another user. After removing the 18 users for whom I could not collect enough pre- or post-treatment tweets (see Appendix A for a full discussion), I used the model trained by Wulczyn, Thain and Dixon (2017) to assign an "aggression score" (between 0 and 1) to each of these 367 thousand tweets. This measure was skewed toward the lower end of the distribution, so I selected all tweets above the 75th percentile aggression score and coded them as incivil.[12]

I selected the 75th percentile based on the empirical distribution of aggression scores; see Appendix B for this distribution. This exact cutoff was not specified in my pre-analysis plan, but the fact that I would be using this model was pre-registered. This kind of pre-registration is especially important when using measures derived from text. Text data is extremely high dimensional, so the development of measures *ex post* allows researchers (often unknowingly) to select the measure out of millions of potential measures that best supports their hypothesis.

To ensure that the Wikipedia model was performing reasonably well in the classification of tweets, I had a random sample of 1,000 subject tweets labeled by crowdworkers on Mechanical Turk. The model correctly predicted the human-generated labels 82% of the time (for a more extensive discussion, see Appendix C).

To control for each subject's pre-treatment behavior, I calculated their rate of incivil tweeting in the three months before the experiment. This measure was included as a covariate in all of the following analysis. I then calculated this same measure for different post-treatment time periods, to test for effect persistence.

[12] A more conservative approach to classifying incivility would set a higher threshold for the aggresion scores. Appendix D presents the results of the same regressions as are presented below, with the threshold set to the 90th percentile of aggression scores. Results are largely unchanged except that there are smaller effects in the 1-day time period. Because the treatment could affective the distribution of aggression scores, I looked only at pre-treatment tweets when calculating these percentiles.

Table 2.1: Distribution of Incivil Subject Tweets, Pre- and Post-Treatment

	1st Quartile	Median	3rd Quartile	Mean
Pre-Treatment (90 days)	125	365	834	579
Pre-Treatment incivil	30	86	221	145
Post-Treatment (43 days)	101	317	773	588
Post-Treatment incivil	25	77	189	137
Republicans: pre-Treatment incivil	30	80	205	126
Democratss: pre-Treatment incivil	31	100	247	165

The data take the form of overdispersed count data: the variables that record the number of incivil tweets sent by each user are bounded by zero and vary widely between highly active and normal users. Table 2.1 reports these distributions. To account for this high variance, I take the log of the incivil tweet count variables in the following results.[13]

The experimental results on the full sample with all treatments pooled are displayed in Figure 2.4; in all of the analysis that follows, the dependent variable is the (log of the) number of incivil tweets the subject sent in the specified time period. Each of the four results are for a given non-overlapping time period; the first result is for the first day after treatment, the second result for days 2-7, the third for days 8-14, and the fourth for days 15-28.

In none of the these four time periods is there a statistically significant treatment effect with all three treatments pooled. To test the second part of Hypothesis 1 (that the effect of the two moral treatments would be larger than the Public treatment), Figure 2.5 pools the three treatments into these two categories. In Week 1, there is a statistically significant effect of the two moral treatments, while the effect of the Public treatment is almost exactly 0. These treatment effects are not, however, statistically significant from each other ($p=.12$).

Although not an explicit hypothesis, it is plausible that treatment effects should decay over time: we should expect to see the largest treatment effects in the the first time period. This does not seem to be the case in Figure 2.5: there are no significant treatment effects

[13]Another approach to handling overdispersed count data is to fit a negative binomial model. The results of this model can be found in Appendix E. Another approach would be to divide subjects by how often they tweeted, to see if treatment effects are constant across subject loquacity. Appendix F shows that this is not the case: treatment effects on the more active subjects are close to zero, while the effects (of the moral treatments) on the less active subjects are significant in several different time periods.

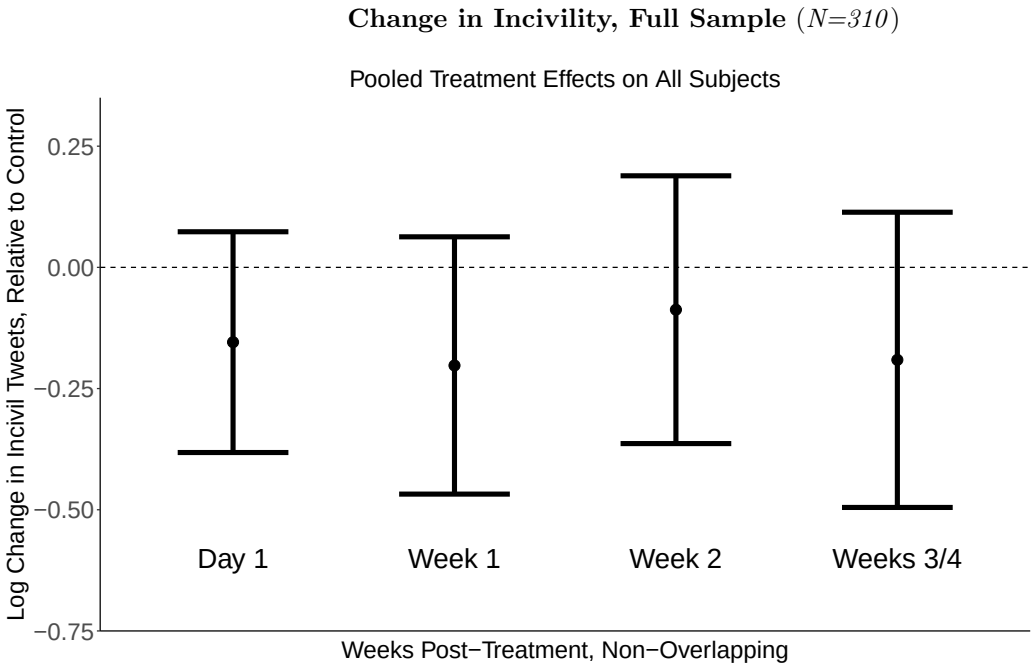


Figure 2.4: Pooled treatment effects on the entire sample, controlling for the log of the number of pre-treatment incivil tweets sent by each subject. Each of the four results are for a given non-overlapping time period; the first result is for the first day after treatment, the second result for days 2-7, the third for days 8-14, and the fourth for days 15-28. Lines represent 95% confidence intervals.

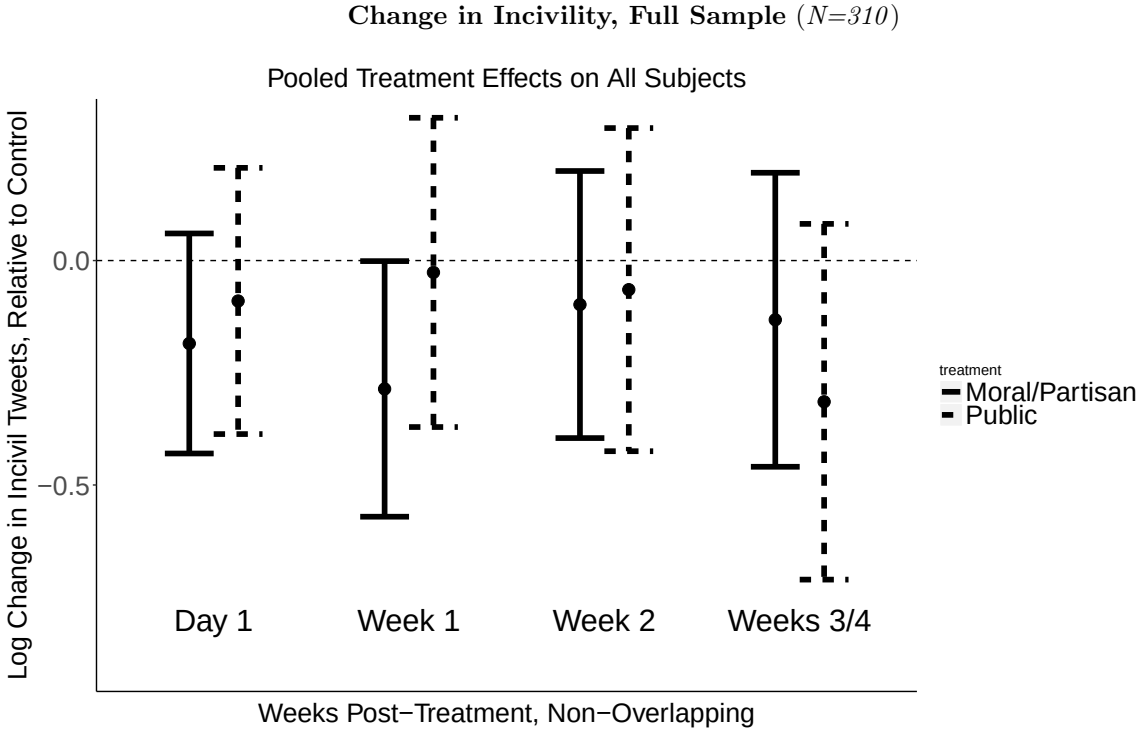


Figure 2.5: Pooled treatment effects on the entire sample, controlling for the log of the number of pre-treatment incivil tweets sent by each subject. Each of the four results are for a given non-overlapping time period; the first result is for the first day after treatment, the second result for days 2-7, the third for days 8-14, and the fourth for days 15-28. Lines represent 95% confidence intervals.

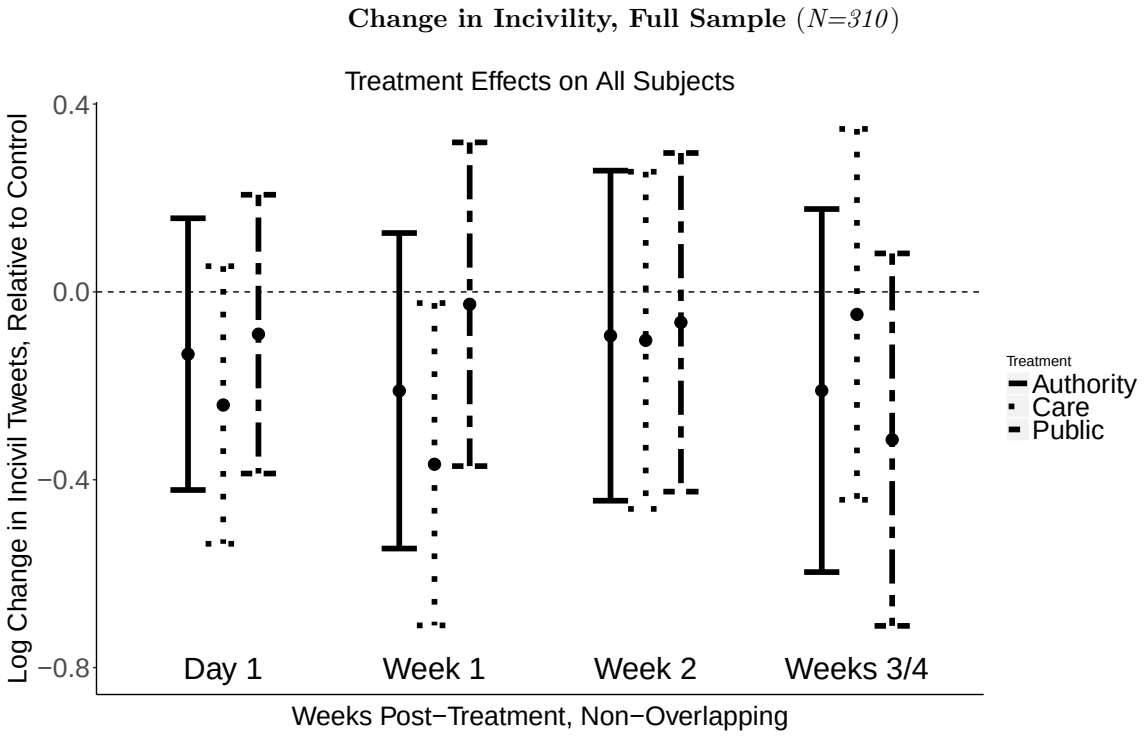


Figure 2.6: Treatment effects on the entire sample, controlling for the log of the number of pre-treatment incivil tweets sent by each subject. Each of the four results are for a given non-overlapping time period; the first result is for the first day after treatment, the second result for days 2-7, the third for days 8-14, and the fourth for days 15-28. Lines represent 95% confidence intervals.

in the first day after treatment, but there are in the first week after treatment.[14] This is to some extent an issue of data size: the first twenty-four hour time period is simply too noisy. The results in Appendix D support this conclusion; if the analysis is run with the higher threshold for classifying tweets as incivil (thus decreasing the number of tweets in the analysis), the point estimate of the effect in Day 1 becomes negligible while the effect becomes larger in Week 1.

Figure 2.6 further disaggregates the treatment effects. In the main Week 1 time period, there is a significant effect of the Care treatment, but not the Authority treatment; these treatment effects are not significantly different from each other, however.

[14]There also appears to be a shift in the effect of the Public treatment: despite negligible effects in the first three time periods, the point estimate shifts down in period four. There is no obvious theoretical justification for this delayed effect, so I believe it is nothing more than (non-significant) noise.

To test the first portion of Hypothesis 1, Figure 2.7 breaks down Figure 2.6 by the partisanship of the subjects. The top panel displays the results for Republicans, and the bottom panel for Democrats.

These results do not support Hypothesis 1. For Democrats, there are no effects in Day 1, while in Week 1, the effects of the two moral treatments are identical and significant (p=*.07*). This is in contrast to the hypothesis that the Care treatment would be more effective than the Authority treatment.

For Republicans, there is a significant effect of the Care treatment in Day 1, but no effects in Week 1 or later. This is in contrast to the hypothesis that the Authority treatment would be more effective than the Care treatment; in fact, the point estimates are in the opposite of the expected direction.

To test Hypothesis 2, I re-ran the analysis in Figure 2.6 with an interaction term between subject anonymity (on the three-point scale, where 0 means they provided a full bio and 2 means they were fully anonymous).[15] Figure 2.8 reports the results for the Day 1 (top panel) and Week 1 (bottom panel) time periods.

In the Day 1 time period, subject anonymity behaves as expected for the two moral treatments: less anonymous subjects change their behavior more by sending fewer incivil tweets. This interaction effect is statistically significant for the Care treatment at $p < .05$, but not for the Authority treatment (p=*.11*).

This trend does not obtain in the Week 1 time period. There is no evidence of any interaction effect for the Authority or Public treatment conditions, and the evidence for such an effect with the Care condition is very weak. There were no significant treatment effects for the later time periods, and similarly no evidence of heterogenous treament effects (results not shown).

One possible explanation for the lack of the expected effect on Democrat subjects is

[15] Given sample size constraints, it is not possible to break down the results simultaneously by subject anonymity, partisanship, and treatment condition; each subject pool would contain approximately 12 subjects.

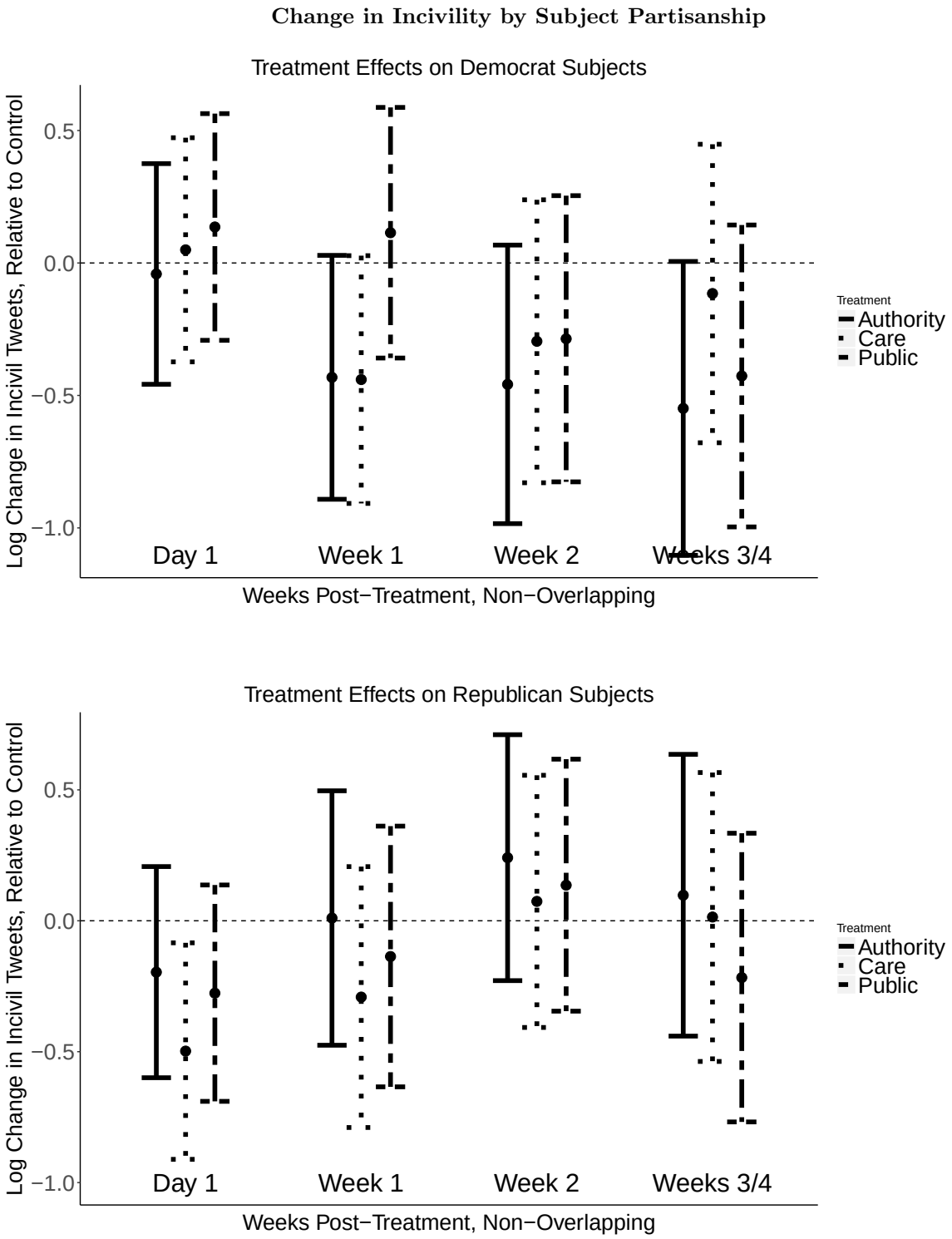


Figure 2.7: Treatment effects divided by subject partisanship. The top panel displays results for Democrat subjects (*N=147*), while the bottom panel displays results for Republican subjects (*N=163*). Lines represent 95% confidence intervals.

Change in Incivility by Subject Anonymity (*N=310*)

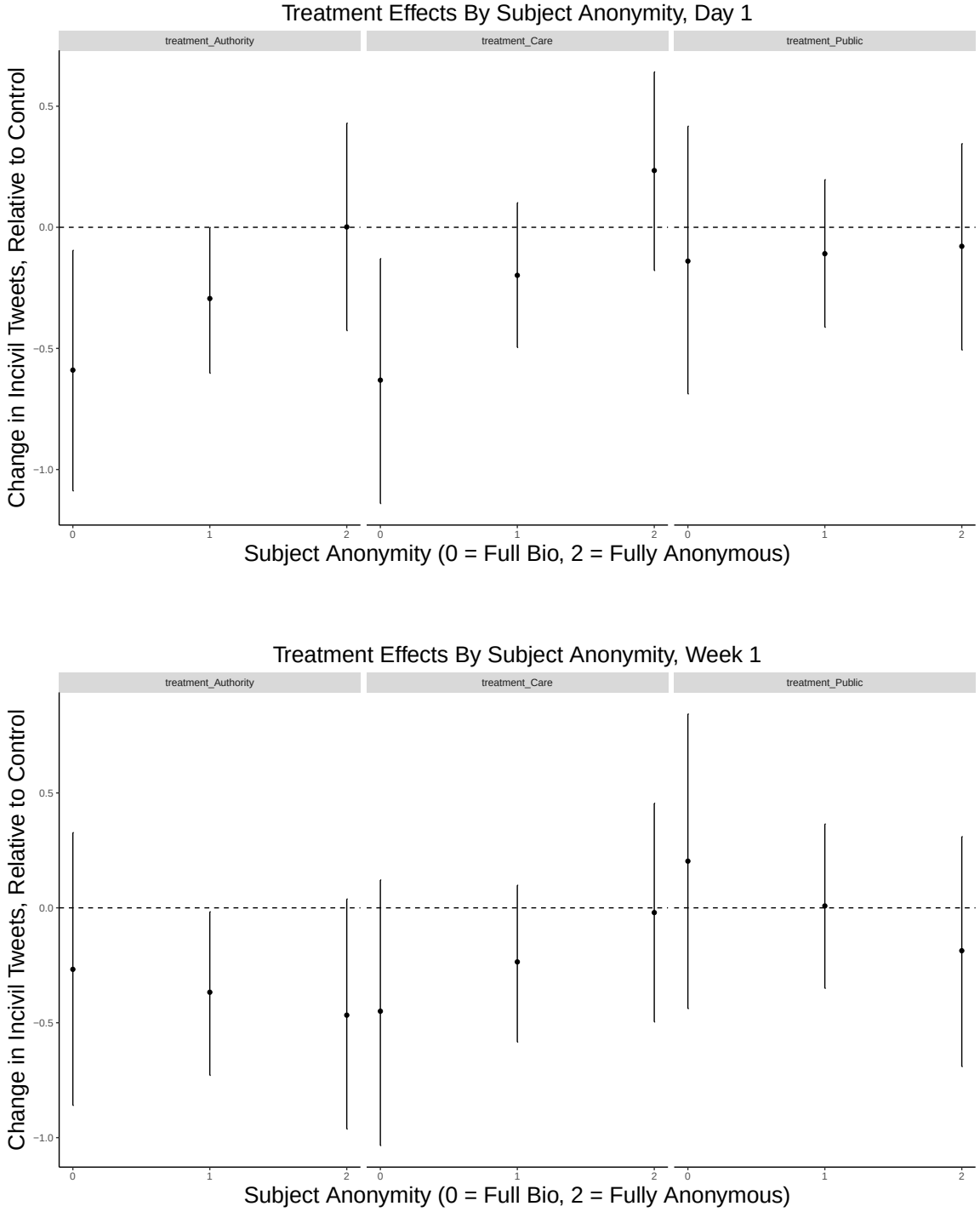
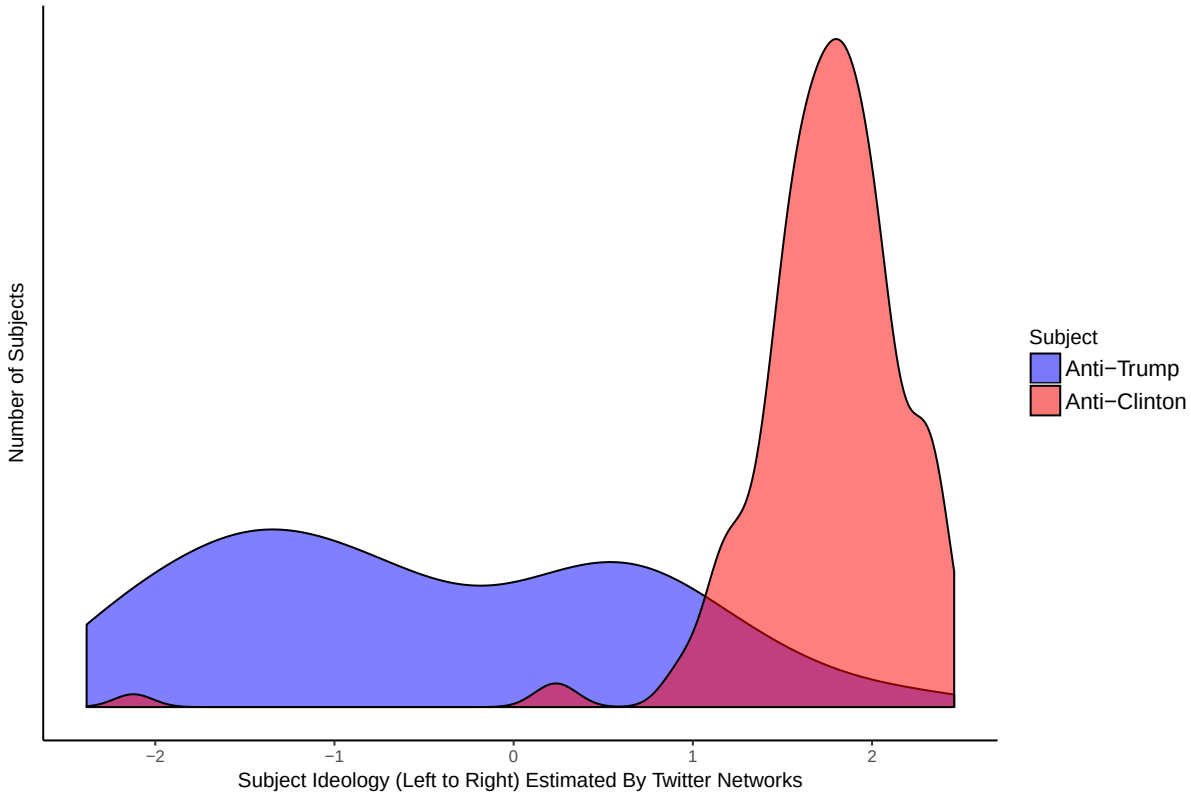


Figure 2.8: Treatment effects divided by subject anonymity. The top panel displays results for the first day after treatment, while the bottom panel displays results for the first week. Lines represent 95% confidence intervals.

Figure 2.9: Estimated Ideology of Subjects Labeled "Republican" or "Democrat"



that this group was more heterogeneous. I implemented the method developed by Barberá (2015) to estimate subjects' ideological ideal points. As Figure 2.9 demonstrates, there was significant heterogeneity in the ideal points of subjects I coded as Democrats, but not for Republicans.

All but two of the subjects coded as Anti-Hillary (Republicans) had estimated ideology scores above 1, and only one was coded as left of center. However, a full third of the subjects coded as Anti-Trump (Democrats) had estimated ideology scores right of center, although only a few are far to the right (have an ideology score above 1). Looking at Figure 2.9, there appears to be two distinct clusters of Anti-Trump subjects; in addition to the expected group of Democrats, there is also a significant contingent of moderate Anti-Trump Republicans that I classified as Democrats. Because the Care and Authority treatment messages were explicitly designed to appeal to subjects' partisan group identities (and identified the Anti-

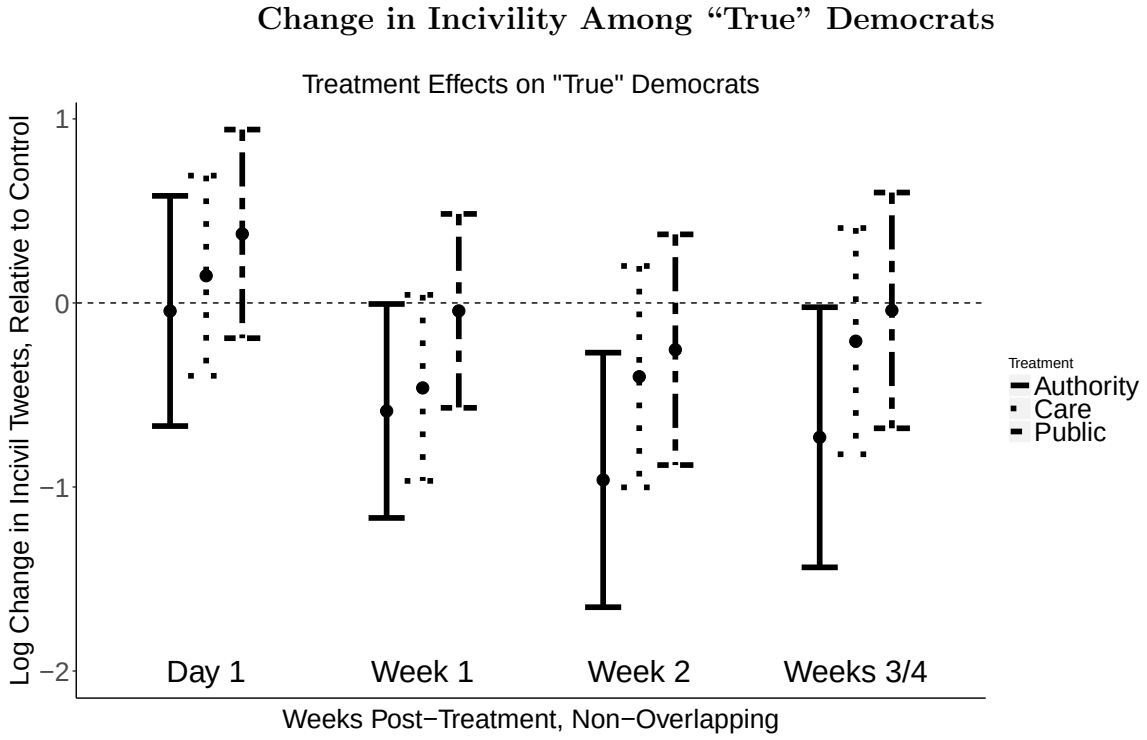


Figure 2.10: Treatment effects on Democrat subjects, restricted to subjects whose ideologies were estimated to be left of center (*N=86*).

Trump subjects as "Democrats"), the ideological heterogeneity within this group could pose a problem for estimating average treatment effects.

If I restrict the analysis of Democrats in Figure 2.15 to only those with estimated ideology scores to the left of center, I find support for this *ex post* explanation. The point estimates for the two moral suasion treatment effects become more negative in the Week 1 and Week 2 time periods, seen in Figure 2.10. Because the sample size is down to 86, the Care treatment is still only significant at *p=.08*, but the largest change is on the Authority effects, which are now significantly negative in the Week 1, Week 2 and Weeks 3/4 time periods.

2.6 Conclusion

The 2016 US Presidential Election took place in the context of a deeply polarized electorate. Many partisans refrain from engaging in political discussion in their day-to-day lives for fear

of alienating members of their communities: Berry and Sobieraj (2013) performed dozens of in-depth interviews with partisans who explained that they often self-censored to "avoid offending others or engaging in awkward social exchanges." However, the authors noticed an asymmetry between liberals and conservatives–"conservative respondents alone...[fear] being judged negatively *as people* because of their view" (emphasis in original).

This offers an explanation for the main (unexpected) result from the experiment discussed in this paper: social sanctioning from Twitter bots was more effective at causing Republicans to decrease their rate of incivil tweeting in the first day after they were sanctioned, but this effect did not persist to the following week. Democrats, however, did not change their behavior immediately, but did reduce their rate of sending incivil tweets in the following week. This asymmetry is important, and drives home the importance of research designs which are able to measure effect persistence. If the analysis is restricted to Democrats whose ideology was estimated to be left of center, the most effect treatment condition caused a significant change in their behavior up to a month later.

I failed to find support for Hypothesis 1: there was little difference between sanctioning language designed to appeal to subjects' moral sense of Care or Authority. My *post-hoc* explanation for this conclusion is that this election was not normal. I expected that a message reminding subjects of the rules of political civility would be more effective on Republicans, but in the 2016 US Presidential election, it was Democrat Hillary Clinton who explicitly positioned herself on the side of civility.

Further, the minimal response from Democrats to the Care treatment may be explained by the tweets they sent to my bots in response to being sanctioned. In several cases, Democrats told my bots something like "these other people are Trump supporters, so I don't care about their feelings"; no Republicans expressed a similar sentiment. This is in keeping with the theory developed in Haidt (2012): the morality of Democrats in the US is based largely on care for specific victim groups, a category which does not include Trump supporters.

Another insight from Berry and Sobieraj (2013)'s partisan interviews is that this re-

straint from talking contentious politics might be context-specific: one subject "[wasn't] rattled by social conflict, as she is comfortable being politically contrarian under the cloak of anonymity." The subjects in my study may have felt similarly: in keeping with my expectations, the treatment effects were largest on the subjects who were the least anonymous, although this trend was only observed in the first day after treatment.

This was somewhat surprising because the subjects' anonymity played a significantly different role in Munger (2017*c*)'s experiment using Twitter bots to sanction users engaged in racist harassment. The role of anonymity in moderating how people engage in online communication is a complicated one, but in the context of Twitter, a semi-anonymous platform in which each user can select her own level of anonymity, these moderating effects are likely to signal differences in the type of user rather than the impact of anonymity *per se*.

My *post-hoc* explanation for the inverted relationship between anonymity and treatment effectiveness in the two studies comes from the composition of the subject pool in each case. Among people using racist slurs, the ones who provided a full biography were fully committed to and unashamed of this behavior, and the treatment was more effective on the more impressionable anonymous users who were aware that their behavior was wrong. The behavior sanctioned in the current study, sending incivil tweets at partisans from the other side, is less objectionable than tweeting racist slurs. The fully anonymous users in this sample, then, may have been more likely to be committed "trolls" than normal (if passionately polarized) people.

This finding fits in with recent research on online trolling, and suggests a way to improve online discourse. Cheng et al. (2017) finds that there are a small number of dedicated online trolls, but that a much larger group of people will use incivil language on forums where others have already been incivil. These are precisely the people who may constitute the subject pool of this experiment: they saw others say something nasty to their preferred candidate, and responded in kind.

It may be difficult to prevent hardcore trolls from setting an incivil tone, but my findings

suggest that it may be possible to prevent incivility from becoming the norm by reminding normal people of our shared humanity and responsibility to the rules of civil discourse.

There is a role for companies like Twitter to set up their platforms to minimize the likelihood of toxic speech. In particular, Twitter's combination of anonymous, easily-created accounts and verified accounts used by journalists and politicians to have serious discussions is an almost ideal playground for trolls foreign and domestic. Restrictions on the activities that can be taken by young accounts could combat this problem, as could imposing tiny costs on accounts that post frequently.

Ultimately, though, political discussions in the physical America are broadly civil; this is not because the government or some other central authority enforces civility, but because of a generally observed norm of civility. Humans have evolved mechanisms for successfully enforcing norms in their real-world societies, but for a variety of technological and cultural reasons, these techniques do not trivially translate to the world of Twitter. The goal of this research is to understand the mechanisms underlying online norm enforcement, to enable people to effectively police their online communities.

The current results are far from sufficient in establishing a universal theory of online norm enforcement, but they point towards a research agenda that can continue to grow our understanding of the topic. Future empirical studies could test other theories of persuasion—collective guilt, appeals to self-interest—and should vary the identities of the bots (race, gender, institutional affiliation). A related research design could employ confederates rather than bots: real human volunteers, embedded in existing networks, could help us understand the network effects of online norm enforcement.

The stakes of improving online political discourse are high: the social web could fulfill the promise of widespread deliberative democracy. If partisan incivility becomes further established as the norm in online communication, it could lead to further affective polarization and self-segregation, creating entirely separate epistemic communities and rendering deliberation impossible.

Appendix for Chapter 2

B.1: Attrition

Although I initially recorded 330 subjects as belonging to either a treatment or control condition, the final analysis includes only 310 subjects. The sample suffered from attrition from one of four sources.

In the case of four subjects, I mis-applied the treatment. When I used my bots to tweet at the subjects, I made a computer error and tweeted directly at them rather than in response to a specific incivil tweet. I became aware of this possibility when one subject responded to my tweet in confusion; in re-checking the rest of the subjects, I found the other 3 mistakes.

I identified the rest of the potentially problematic subjects through patterns in their tweeting behavior. I manually re-inspected all of the profiles of subjects for whom I collected fewer than 50 tweets pre-treatment *and* 50 tweets post-treatment. The majority of the profiles I identified this way still merited inclusion; they were just people who did not tweet very often. However, I excluded others from the final sample. I did this manual re-inspection before calculating any of the results and without knowledge of the treatment condition to which the subjects belonged.

The most common problem was that I had 0 pre-treatment tweets for a subject despite having thousands of post-treatment tweets. This was caused by the timing of when I scraped their profiles and the Twitter API's historical tweet limit: Twitter will only give you the 3,200 most recent tweets from a given account. I performed a full scrape of each account within a week of the treatment; this implies that these accounts were tweeting thousands of times a week. This is very difficult for a human to do, so I suspect that many of these accounts were bots; if they were not bots, they were extremely atypical Twitter users. However, this was the single largest source of attrition; just under 3% of the original accounts were excluded for this reason.

There were a total of 3 accounts in my sample that were suspended by Twitter during

Table 2.2: Attrition Rates and Causes

	Control	Democrats	Republicans
Initial assignment	108	104	118
Failed treatment application	0	2	2
Tweeted too often/bots	3	1	5
Suspended	0	1	2
Weird	2	0	0
Final	102	100	108
Attrition	6%	4%	8%

the course of my experiment. I do technically have enough tweets from these accounts to include them in the analysis, but doing so has the potential to bias my results upwards: the reduction in the number of incivil tweets they sent was actually caused by Twitter preventing them from tweeting, rather than by the treatment.

Finally, there were two accounts that were just weird; they had not tweeted thousands of times, but each still only recorded 3 pre-treatment tweets. In both cases, the accounts appeared to be behaving very oddly, and since I did not have a reasonable estimate of their pre-treatment behavior, I excluded them.

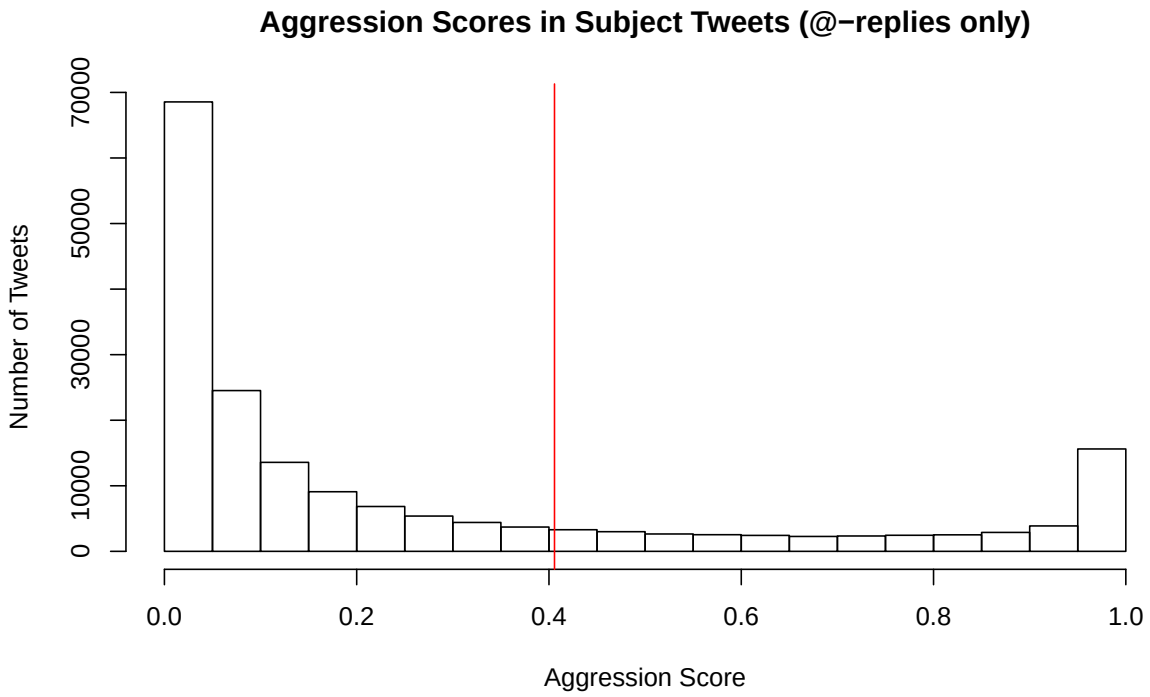


Figure 2.11: Empirical distribution of aggression scores. The vertical line represents the 75th percentile, the cutoff I use in the body of the paper.

B.2: Empirical distribution of aggression scores in subject tweets

As shown in Figure 2.11, the distribution of aggression scores (as coded by the algorithm developed by Wulczyn, Thain and Dixon (2017)) is bimodal: there is a large cluster of "non-aggressive" tweets near 0, and a smaller cluster of "definitely aggressive" tweets near 1. The vertical line represents the 75th percentile of this empirical distribution, the cutoff I use in the body of the paper for transforming these scores into a binary measure. The higher cutoff of the 90th percentile would entail including only the far-right cluster of tweets; the main results are replicated using this higher threshold in Appendix D.

B.3: Validation of Wikipedia measure on the current dataset

Figure 2.12 plots the accuracy of the scores derived from the Wulczyn, Thain and Dixon (2017) model in predicting the labels of tweets coded by crowdworkers. The x-axis plots the threshold used to turn the continuous scores output by the model into binary labels; there is a slight peak (accuracy = .82) at the vertical line, which depicts the 75th percentile used in main results from the text, but the accuracy is fairly constant across a wide range of cutoffs.

The validation tweets consist of 1,000 tweets which were randomly sampled from among all subject tweets and uploaded to Mechanical Turk. Each tweet was coded by two of Amazon's "Expert Coders," a restrictive label that they only award to consistently attentive crowdworkers. The precise instructions given to the workers were as follows:

```
Please read each tweet and tell us if it is civil or incivil.

We say that "civil" tweets are those that demonstrate respect for the
person being tweeted at.

If a tweet has very little information (if it just contains a link, for
example), code it as "civil."
```

Overall levels of intercoder reliability were low by the standards of objective classification tasks (Krippendorf's alpha = .37). The task at hand, however, is inherently subjective, and our results are in line relevant published work: Wulczyn, Thain and Dixon (2017), using a somewhat more rigorous coder vetting process, report "a Krippendorf's alpha score of 0.45. This result is in-line with results achieved in other crowdsourced studies of toxic behavior in online communities. (p3)"

For the accuracy results displayed in Figure 2.12, I restricted the initial 1,000 tweets to the 70% on which the coders agreed on the label. These labels are unbalanced in the sample

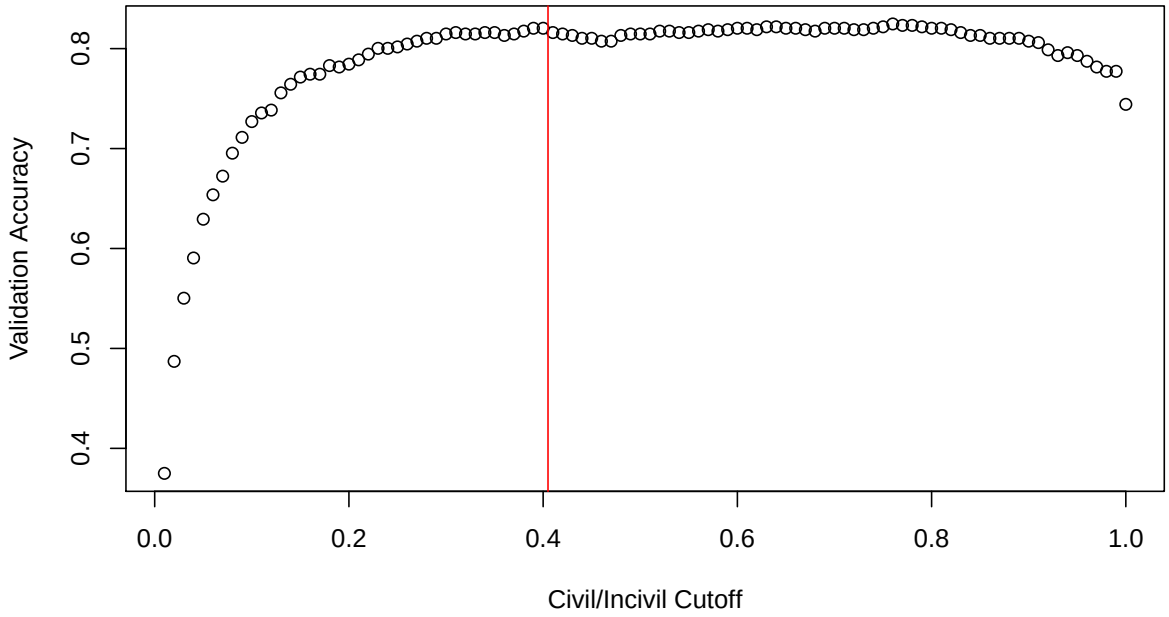


Figure 2.12: Accuracy of the Wikipedia model applied to tweets labeled by Mechanical Turk workers, scored on the tweets on which coders agreed on whether the tweet should be labeled civil or incivil. The vertical line represents the 75th percentile, the cutoff I use in the body of the paper.

(74% were labeled civil), so the 82% accuracy represents a significant improvement on a naive classification scheme.

B.4: Main Results Using Higher Aggression Threshold

The results in the body of the paper use the 75th percentile of the empirical distribution of aggression scores (as coded by the algorithm developed by Wulczyn, Thain and Dixon (2017)) to code subject tweets as civil or incivil. There is a distinct cluster of "definitely aggressive" tweets near the top of this distribution, and the results in Figure 2.13 plot the model results when only this cluster is coded as incivil—that is, using the 90th percentile as the threshold. In both plots, the effects in the 1 Day time period become closer to 0, while the effects in the 1 Week time period become more pronounced.

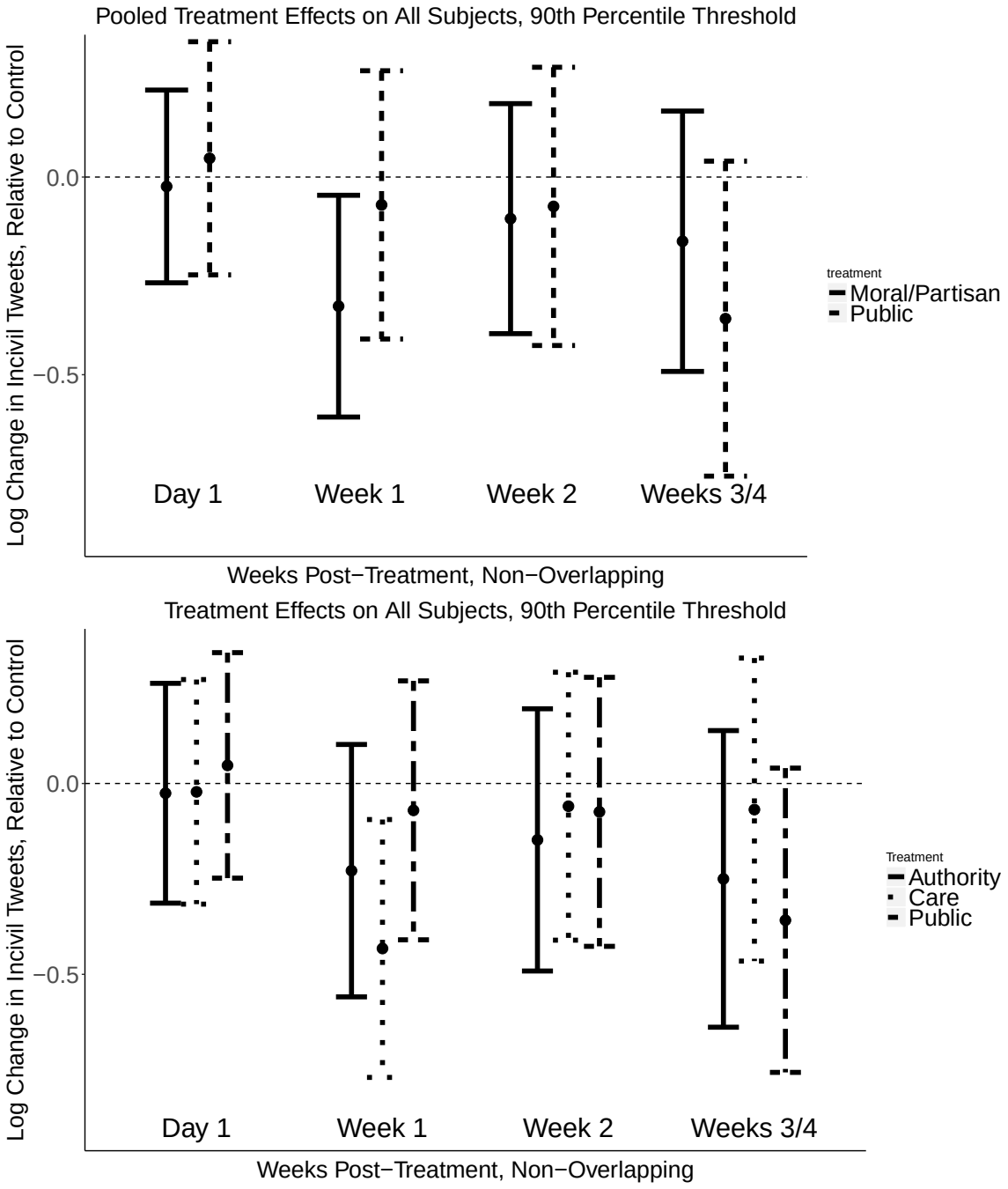


Figure 2.13: Main results replicated using the higher threshold of aggression scores for coding tweets as incivil.

B.5: Negative Binomial Specification of Main Results

The dependent variable of interest in this analysis is the number of times a subject sent an incivil tweet to another user. This is a "count variable"–it can only take non-negative integer values–and thus violates a fundamental assumption of OLS regression. To address this issue, generalized linear models with different assumptions are often used. Poisson regression, in which the dependent variable is assumed to have a Poisson distribution, is a common technique, but this carries the further assumption that the variance and expected value of the dependent variable are equal. In cases in which the variance is significantly higher than the expected value–like it is here–the negative binomial model relaxes this assumption (Hilbe, 2008).

$$ln(Agg_{post}) = x_{int} + \beta_1 Agg_{pre} + \beta_2 T_{feel} + \beta_3 T_{rules} + \beta_4 T_{public} + \beta_5 Anon + \beta_6 (T_{feel} \times Anon)$$

$$+\beta_7 (T_{rules} \times Anon) + \beta_8 (T_{public} \times Anon)$$

To interpret the relevant treatment effects implied by the coefficients estimated by this model, the exponent of the estimated $\hat{\beta}_k$ for each of the treatment conditions needs to be added to the corresponding $\hat{\beta}$ for the interaction term, evaluated at each level of Anonymity Score (Hilbe, 2008). For example, the effect of the Feelings treatment on subjects with Anonymity Score 1 (the middle category) is:

$$IRR_{feel \times Anon_1} = e^{\hat{\beta}_2 + \hat{\beta}_6 \times 1}$$

The results of these negative binomial models can be seen in Figure 2.14 and Figure 2.15.

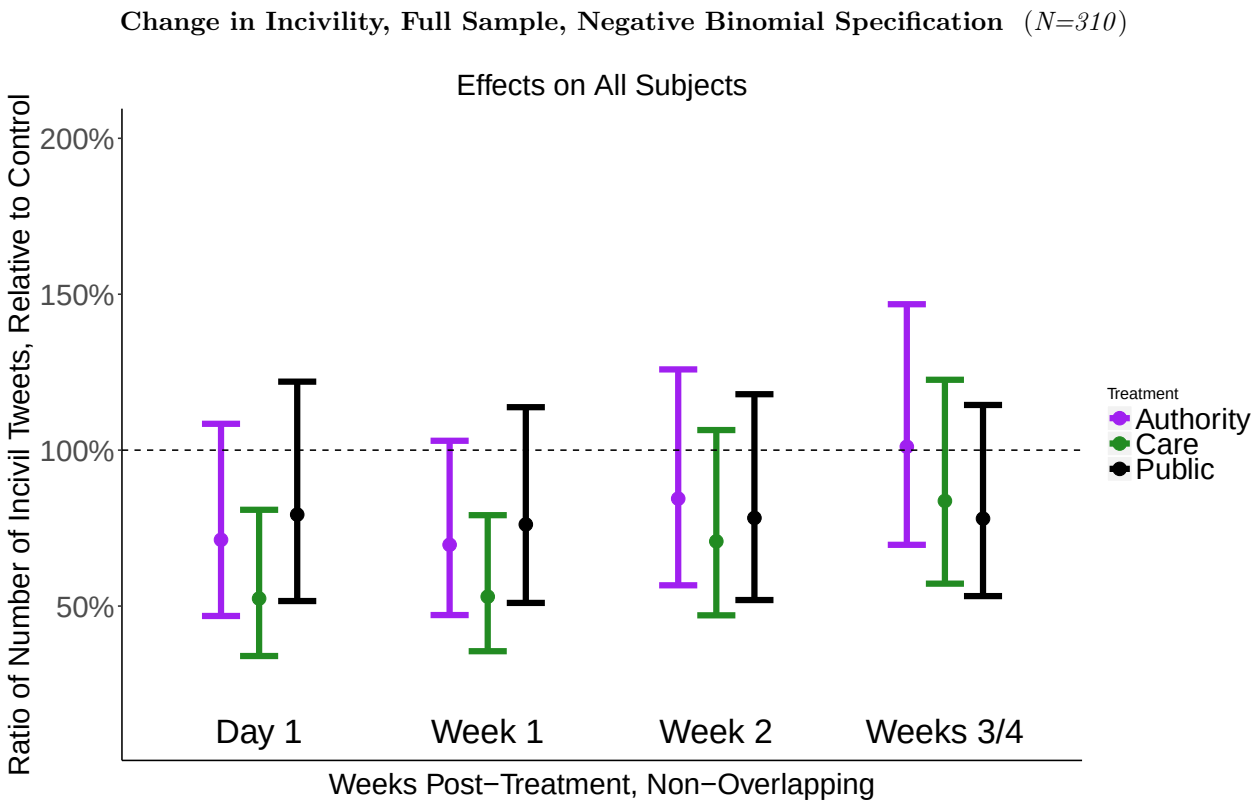
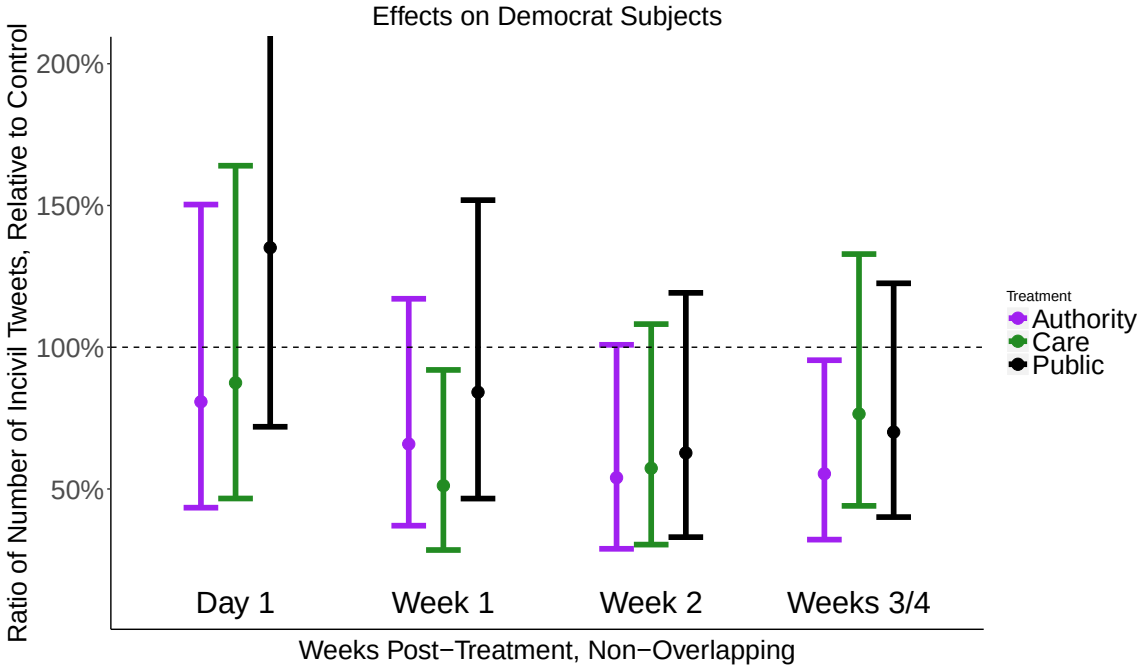
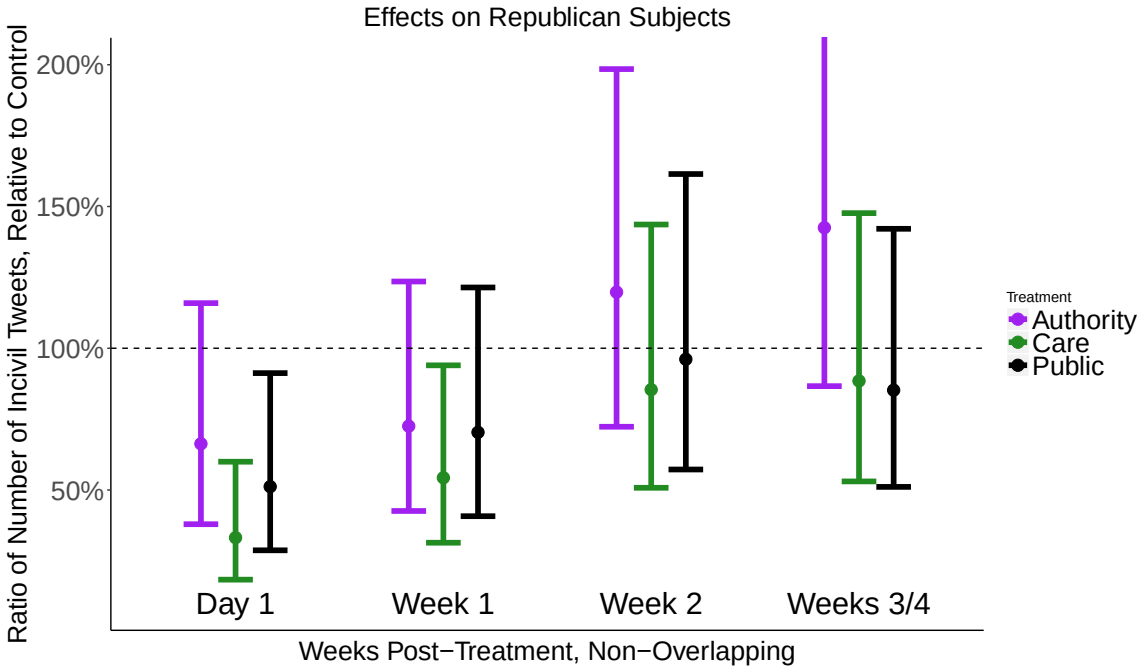


Figure 2.14: The Incidence Ratio calculated from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot means that these subjects sent 50% as many directed incivil tweets as the subjects in the control group. 95% confidence intervals.

Figure 2.15: **Effects on Democrats, Negative Binomial Specification** (*N=147*)



Effects on Republicans, Negative Binomial Specification (*N=163*)



The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot in Panel A means that these subjects sent 90% as many directed incivil tweets as the subjects in the control group. 95% confidence intervals.

B.6: Results Divided by Subject Loquacity

The subject population is highly varied in their pre-treatment level of tweeting activity. Although not one of the tests I specified in my pre-analysis plan, the potential policy implications of heterogeneous effects based on subject loquacity merit investigation of this possibility.

Figure 2.16 replicates the main results in the paper by the pre-treatment tweeting rate of the subjects. The top panel displays the results for subjects above the median (82 incivil pre-treatment tweets), and the bottom panel for subjects blow this threshold.

There is a clear distinction: treatment effects on the more active subjects are close to zero, while the effects (of the moral treatments) on the less active subjects are significant in several different time periods.

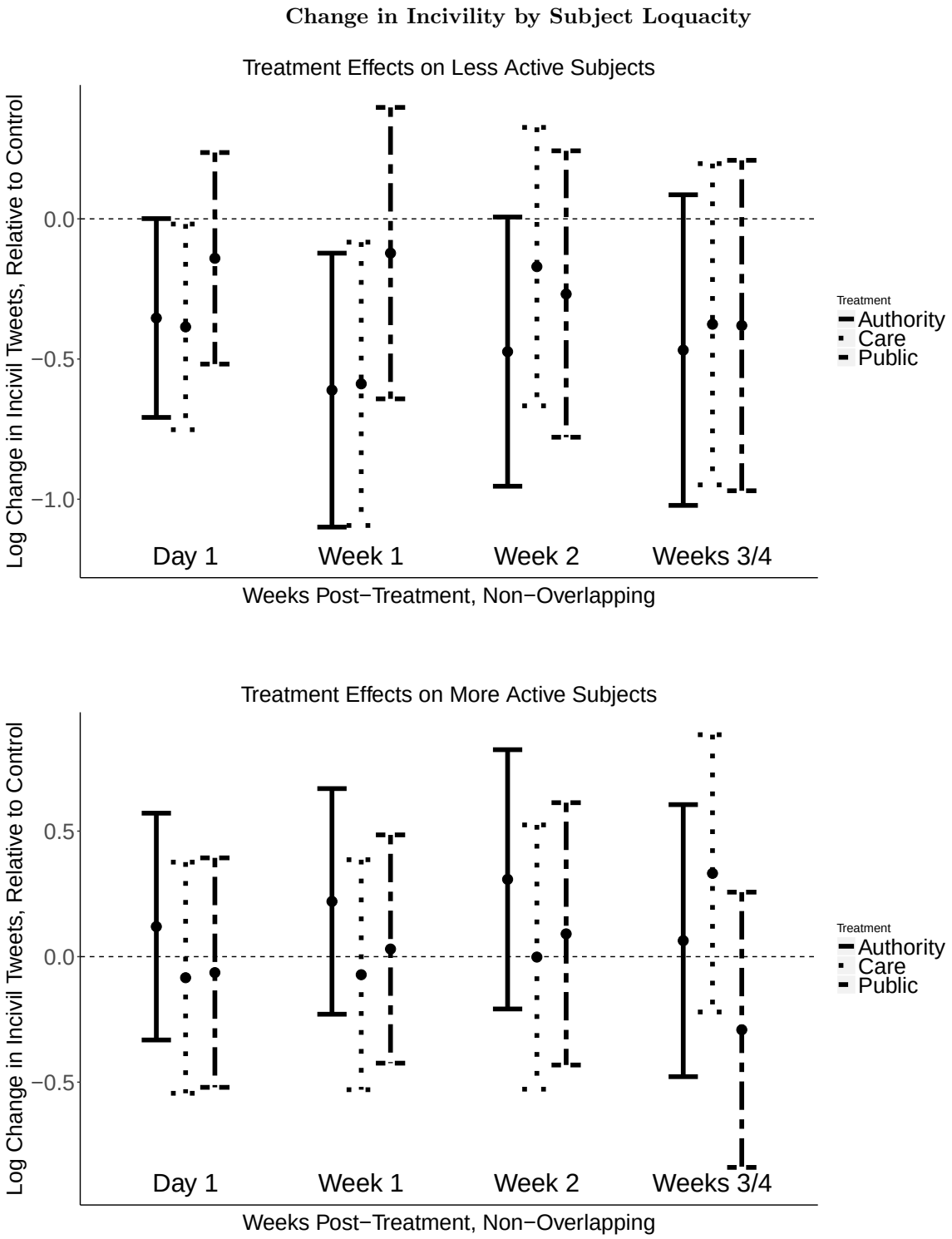


Figure 2.16: Treatment effects divided by subject pre-treatment tweeting rate. The top panel displays results for less active subjects (below the median), while the bottom panel displays results for more active subjects (above the median). Lines represent 95% confidence intervals.

B.7: Treatment Effects on Sending Civil Tweets

The results in the body of the paper display treatment effects on the rate of sending *incivil* tweets; it is worth exploring whether the treatment had an analogous effect on sending *civil* tweets.

I re-ran the analysis using the number of *civil* tweets as the dependent variable (those with aggression scores below the 75th percentile threshold), and found no significant treatment effects. The point estimates are in the same direction as the effects on incivil tweets, with effect sizes ranging from 50% to 80% as large. Figure 2.17 displays these results. In Panel A, examining pooled treatment effects, these effect sizes are for the civil tweets, .08 (1 Day) and .16 (1 Week); for the incivil tweets in the body of the text, these effect sizes are, .15 and .2.

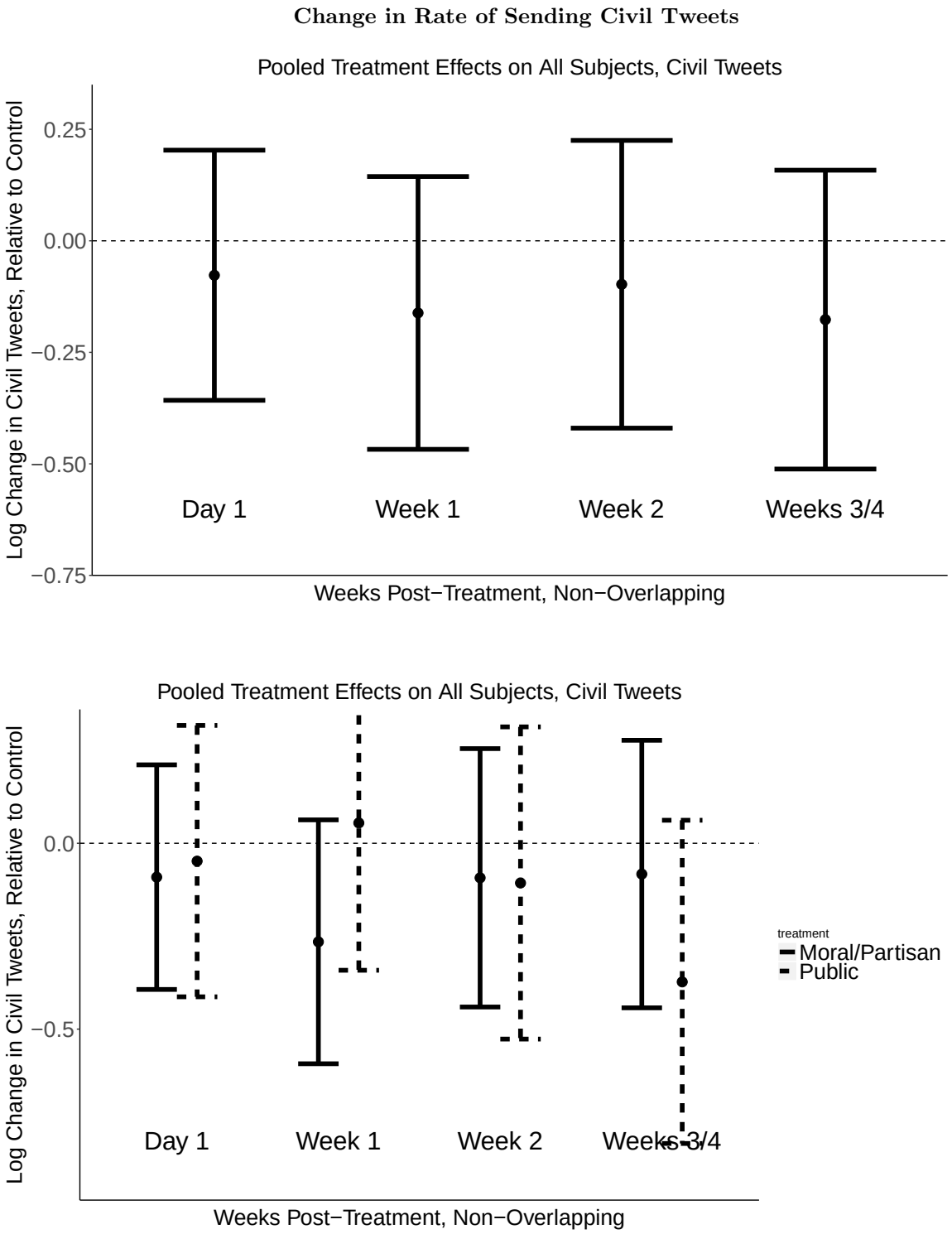


Figure 2.17: Treatment effects on the rate of subjects sending civil tweets (tweets scored as below the threshold for incivility used the body of the paper). The top panel displays results pooled across all treatment conditions, while the bottom panel displays results where the two moral treatments are pooled. Lines represent 95% confidence intervals.

Chapter 3

The Effect of Clickbait

In the Introduction, I outlined my theory of Clickbait Media and how this paradigm describes the structure of the online media industry today. A crucial portion of this theory involves a specific format of news article that is commonly called "clickbait"; to repeat the dictionary definition, this is "something (such as a headline) designed to make readers want to click on a hyperlink especially when the link leads to content of dubious value or interest."

This conception of clickbait thus has a negative connotation, characterized by something like regret—if a consumer of clickbait stopped to think about their decision to click on such a story, they would probably not do so.

I do not think this definition is up to date; as discussed in the Introduction, Facebook was able to detect this form of deceptive clickbait (by looking for instances in which users clicked a link and then quickly closed it) and penalize internet media firms that employed it.

A more concerning form of clickbait is one that appeals directly to people's fears, especially as it relates to a threat to a social group to which they belong. This form of clickbait serves the twin purposes of inducing excitement by appealing to group competition (Abramowitz and Saunders, 2006) and being easily spread among online social networks, which tend to be homophilous. Although not all emotional clickbait is low-quality, the *Credibility Cascades* described in the Introduction can serve to spread low-quality news stories among a group

that shares a social identity threatened by the news in the story.

Emotional clickbait—when the news is about politics or the relevant social groups are politically relevant—is also more concerning to political scientists than the straightforwardly deceptive information gap clickbait headline because of its capacity to polarize and create separate epistemic communities. In the American context, these concerns manifest themselves as affective polarization between Republicans and Democrats (Iyengar, Sood and Lelkes, 2012) and the potential for filter bubbles (Flaxman, Goel and Rao, 2016). There is the additional concern that all forms of clickbait erode public trust in journalism. Since the heydey of broadcast journalism, he news media has consistently been shifting away from hard news in order to meet audience demand; that audience then has less trust in news media (Ladd, 2011). On the other hand, this eroding trust may also increase political knowledge: previous evidence has shown precisely this trend when comparing civil and uncivil cable news (Mutz, 2015).

This Chapter describes the results of pair of experiments designed to test these possibilities. I first provide evidence that certain types of people are more likely to select an emotional clickbait headline when given the opportunity; this heterogeneity is essential to the theory of Credibility Cascades. Although I did not have a strong theoretical expectation *ex ante*, there is robust evidence that older people, moderate Republicans and more frequent social media users and consumers of online news have a higher preference for clickbait.

I did hypothesize (per-registered through EGAP, number 3175) that random assignment to read stories with emotional clickbait headlines would exacerbate affective polarization, decrease trust in online news, and increase information retention; in every case, I observed null effects. Below, I discuss possible explanations for these null effects and propose new research designs which could help determine the robustness of these null effects.

3.1 Experiments on the Determinants and Effects of Clickbait News Consumption

I conducted two related survey experiments using Amazon's Mechanical Turk. Each experiment was designed to take around ten minutes to complete, and contained an attention check and built-in delays to discourage respondents from giving low-quality answers; I removed any respondents who failed the check. Because the pool of MTurk workers contains more Democrats than Republicans, I supplemented the first draw with a sample of self-reported Conservatives.

The two experiments were similar in design. In each case, respondents were directed to a Qualtrics survey in which they first reported demographic information, including partisan affiliation and their frequency of internet/Twitter/Facebook use. They were then given a series of eight tasks in which they were asked which of four headlines they would most like to read. In each case, there were two political stories (one Democrat-leaning, one Republican-leaning) and two non-political stories (one sports, one entertainment).[1] One of the two political stories in each decision set was turned into a clickbait headline through the addition of an attention-grabbing phrase to the beginning, so that there were four instances in which the Democrat-leaning headline was clickbait and four instances in which the Republican-leaning was clickbait.[2]

With this process, I calculated the individual-level *preference for clickbait*: whether respondents would select the clickbait headlines rather than the non-clickbait headlines, controlling for the preference for non-political headlines.[3]

Respondents were then randomly assigned to one of four treatment conditions (Experiment 2 added a fifth "placebo" condition in which respondents were given a story about

[1] The inclusion of non-political stories has been shown by Arceneaux and Johnson (2013) to provide more reliable estimates of media choice.

[2] To examine the entire survey instrument, see Appendix C.1.

[3] This process balances the concern expressed in Leeper (2016) for measuring media treatment effects on the relevant population (those who would actually consume the given media) with the fact that we needed to disguise the nature of the manipulation from the respondent.

Table 3.1: Treatment Headlines

Experiment 1	
Republican: Baseline	Republican bill doesn't eliminate Obamacare
Republican: Clickbait	This will make your blood boil: Republican bill doesn't eliminate Obamacare
Democrat: Baseline	Republican bill destroys Obamacare
Democrat: Clickbait	This will make your blood boil: Republican bill destroys Obamacare
Experiment 2	
Republican: Baseline	Trump economic policies working
Republican: Clickbait	Democrats won't like this economic news: Trump policies working!
Democrat: Baseline	Trump economic policies not working
Democrat: Clickbait	Republicans won't like this economic news: Trump policies not working!

sports) through a 2x2 treatment design that varied the partisan leaning of a headline and whether the headline was clickbait. In each case, respondents were presented with a hyperlinked headline; when they clicked the headline, they were directed to a separate tab which displayed the given headline and a news story; the text of the news story was held constant across the conditions.

After respondents read the story and closed the tab, they were asked a feeling thermometer question about Republicans, Democrats, online media and traditional media, as well as a multiple-choice question about their trust in online media and traditional media. On the next page, they were asked three (with an additional, placebo factual question in Experiment 2) multiple-choice questions based on facts presented in story they had been given to read.

The difference between Experiment 1 and Experiment 2 stems from the subject matter of the news story and the way that clickbait and partisanship interacted in the construction of the treatment headlines. The story in Experiment 1 was about the Republican health care bill (relevant in October 2017, when the experiment was fielded), while the story in Experiment 2 summarized the findings from the October jobs report. The text of the stories in both experiments was taken from politically neutral news sources: CNN Money and Quartz Media.

The treatment headlines for the two experiments are displayed in Table 3.1.

Experiment 1 (and the analysis I conducted) was primarily exploratory. This fact is reflected in the theoretical confusion evinced in the treatment headline for Experiment 1. The

distinction between the Democrat and Republican base headlines is insufficiently symmetric.

The intention was to design headlines that would anger the respective partisan groups, then amplify that anger through the emotional clickbait introduction. The literature on affect polarization is still being developed, but a well-established trend is that the gap in partisan affect is driven by decreased evaluations of the out-party. This is what Abramowitz and Webster (2016) call "negative partisanship"—out-partisan animosity is a powerful motivator for a range of political behaviors. Mason (2016) finds experimental support for the presence of anger in response to partisan threats.

These results provided the general motivation for my experimental manipulations, but they were not operationalized correctly. The issue stems from the use of the phrase "destroys Obamacare" compared to the phrase "doesn't eliminate Obamacare". These were both meant to represent baseline partisan headlines, with the addition of the phrase "This will make your blood boil" intended to provide the *emotional clickbait*. The problem is that the phrases "destroys" and "doesn't eliminate" are not symmetric in their degree of affect. Further, the audience implied by the phrase "doesn't eliminate Obamacare" is a partisan one, a fact which could confuse respondents from the opposite party randomly assigned to that headline.

To address these flaws, Experiment 2 uses symmetric headlines, adding only a "not" to switch the partisan leaning. The emotion appealed to in this version of emotional clickbait is even more explicitly negative partisan excitement: the idea that your opponents being angry about something implies that you will excited by it (Abramowitz and Saunders, 2006).

The first experiment was conducted on 927 respondents, 780 of which passed the attention check: 268 Republicans, 288 Democrats, and 224 Independents. The second experiment was conducted on 747 respondents, 591 of which passed the attention check: 211 Republicans, 217 Democrats, and 163 Independents. The average time to complete the survey was 9.5 minutes, just under the advertised ten minutes. Each respondent was paid $1.

3.2 Hypotheses

The first question this study hopes to answer is exploratory: what kinds of people are more likely to consume *emotional clickbait?* There is not any strong theory here, so the analysis related to this research question will be descriptive rather than confirmatory.

Research Question: *What kinds of people are more likely to consume emotional clickbait?*

The hypotheses about the effects of being randomly assigned to the clickbait treatment conditions are the same across the two Experiments. After Experiment 1, I explored the data in an exploratory fashion, testing the hypotheses as I saw fit.

The R file containing all of the code used to analyze Experiment 2 was included with my EGAP pre-registration (number 3175) as my pre-analysis plan. In addition to pre-registering the hypotheses listed below, I specified in advance the precise coding and data manipulation decisions I would make in testing those hypotheses. The specific language of the hypotheses has been changed to match the terminology in the rest of this paper.

Hypothesis 1 *The Republican conditions will increase reported affect toward Republicans. The Democrat conditions will decrease reported affect toward Republicans.*

In both Experiments, the Republicans or their president are the primary political actor mentioned in the headline. Hypothesis 1 thus predicts a change in the way that subjects feel about the Republican party, in the direction of the frame of that headline.

Hypothesis 2 *The effects predicted in Hypothesis 1 will be larger in the clickbait than the baseline conditions.*

Hypothesis 2 predicts that the addition of the *emotional clickbait* language to the beginning of the partisan headlines will amplify the effects of those frames.

Hypothesis 3 *Clickbait conditions will be associated with a distrust in (online) media.*

Following the findings in Mutz (2015) on the impact of incivil cable news, Hypothesis 3 predicts that respondents who are assigned to click on a story with a clickbait headline will decrease their trust in media (either just online media or both online and offline media), but Hypothesis 4 predicts that this will be accompanied by an increase in respondents' ability to recall specific facts from that news story.

Hypothesis 4 *Clickbait conditions will be associated with higher information retention.*

3.3 Results

To analyze the individual-level preference for clickbait (PfC), I combine data from Experiments 1 and 2; this portion of the survey came at the beginning and was held constant. This quantity is calculated by estimating what percentage of the political stories the subjects selected to read were clickbait. I also estimate the individual-level preference for Republican (PfR) news. Note that these two quantities are structurally (negatively) correlated: an individual who selected 8 out of 8 clickbait stories would necessarily have selected 4 out of 8 Republican stories.

Partisans made the expected choices: the mean PfR was .64 for Republicans and .34 for Democrats, including leaners. Restricted to strong partisans, these numbers are .32 and .68. This tendency restricts the range of possible values for PfC.

Still, the overall results were surprising: there actually appears to have been an *aversion* to clickbait. The rate of selecting the clickbait political stories was slightly lower than the non-clickbait political stories (median PfC = .50, mean PfC = .46).

In addition to the restricted range discussed above, this result illuminates a limitation of the current research design: respondents were acutely aware that they were taking part in a study, and they may have made less impulsive choice than they would have in a more

naturalistic setting.[4]

With this caveat, though, I can still estimate the subjects' *relative* PfC. Table 3.2, columns 1 and 2, display the results of an OLS regression taking PfC as the dependent variable and all of the demographic information collected from users as independent variables. Questions about frequency of Facebook use, Twitter use, Internet use, reading online news stories and reading offline news stories were on an 8-point scale; for efficiency, the current variables are binary, taking the value 1 if the respondent selected one of the top three categories (indicating that they use a given platform at least once a day). Education levels are similarly binarized, with the variable College Education taking the value 1 if the respondent finished college.

Overall, frequent Facebook or Twitter users have a significantly higher PfC, as do older individuals. Column 2 adds variables that explicitly ask about news consumption habits. Subjects who report reading online news at least once a day have a *much* higher PfC; the addition of this variable reduces the effect size of social media use variables, although the effect of frequent Facebook use is still positive and significant. The most striking result is in the non-monotonicity of the effect of partisan ID on PfC: excluding leaners, the effect of being a Republican or Democrat (relative to the base category of Independent) is almost exactly 0. However, looking only at leaners, there are significant effects: Democrat leaners have a significantly ($p < .10$) *lower* PfC, while Republican leaners have a significantly ($p < .001$) *higher* PfC.

The explanation for this non-monotonicity can be found in column 3. Partisan ID has the expected results on preference for Republican (PfR) stories: as reported PID moves from strong Democrat to strong Republican, the PfR increases monotonically. Because the PfR is so strong among non-leaner Democrats and Republicans, these respondents have no "degrees of freedom" left to choose between clickbait and non-clickbait stories. Notice that none of the media use variables have an effect on PfR—and that the addition of the news

[4]Furthermore, I do not take these results as evidence that clickbait "doesn't work". Dozens of competing media firms have in effect demonstrated that clickbait does work by adopting it as a prominent format for news headlines, sometimes using explicit A/B testing. There does not yet exist reliable descriptive estimates of the prominence of clickbait relative to traditional headlines, however.

Table 3.2: Preference for Clickbait

	Dependent variable:			
	Preference for Clickbait		Preference for Republican	
	(1)	(2)	(3)	(4)
Frequent FB user	0.024**	0.020*	0.010	0.010
	(0.012)	(0.012)	(0.015)	(0.015)
Frequent Twitter user	0.023**	0.017	0.001	0.003
	(0.011)	(0.011)	(0.014)	(0.014)
Frequent internet user	−0.010	−0.024	0.014	0.020
	(0.054)	(0.054)	(0.068)	(0.069)
Age	0.001***	0.001**	0.002***	0.002***
	(0.0004)	(0.0004)	(0.001)	(0.001)
College Education	−0.008	−0.011	−0.006	−0.006
	(0.011)	(0.011)	(0.014)	(0.014)
Frequent offline news		−0.005		0.006
		(0.014)		(0.018)
Frequent online news		0.049***		−0.016
		(0.014)		(0.018)
Democrat	−0.004	−0.005	−0.163***	−0.163***
	(0.015)	(0.015)	(0.019)	(0.019)
Lean Democrat	−0.030*	−0.032*	−0.096***	−0.096***
	(0.017)	(0.017)	(0.022)	(0.022)
Lean Republican	0.061***	0.062***	0.116***	0.116***
	(0.016)	(0.016)	(0.020)	(0.020)
Republican	0.002	0.003	0.182***	0.182***
	(0.016)	(0.016)	(0.021)	(0.021)
Constant	0.402***	0.391***	0.398***	0.401***
	(0.056)	(0.056)	(0.072)	(0.072)
Observations	1,356	1,356	1,356	1,356
R^2	0.033	0.042	0.225	0.226
Adjusted R^2	0.027	0.034	0.220	0.219
Note:	*p<0.1; **p<0.05; ***p<0.01			

consumption variables in column 4 has exactly 0 effect on any point estimates—but that older respondents do have a significantly higher PfR, above and beyond PID.

Turning to the results from the experimental condition, there is little evidence of an effect of clickbait headlines on any of the theorized outcomes. Figure 3.1 displays the results for the feeling thermometers of the different parties, for both Experiments 1 and 2. The results for Experiment 1 are in the left column, Experiment 2 on the right.

In all of Figure 3.1, almost none of the estimated differences in partisan affect are significant across any of the treatment conditions. The one exception, in Experiment 2, is the placebo condition: Republican affect among Republican respondents was significantly lower when they read a story about the NBA than any of the treatment conditions. This effect was large enough to drive an analogous effect on affective polarization among Republican subjects.

The mechanism here is almost certainly the presence of the phrase "Trump's economic policies" in all four of the treatment headlines. This partisan cue explains the increased rating of Republicans relative to a neutral baseline. This finding provides evidence that subjects were in fact responsive to different stories they were assigned to read, and that the lack of the hypothesized effects among the treatment conditions is genuine.

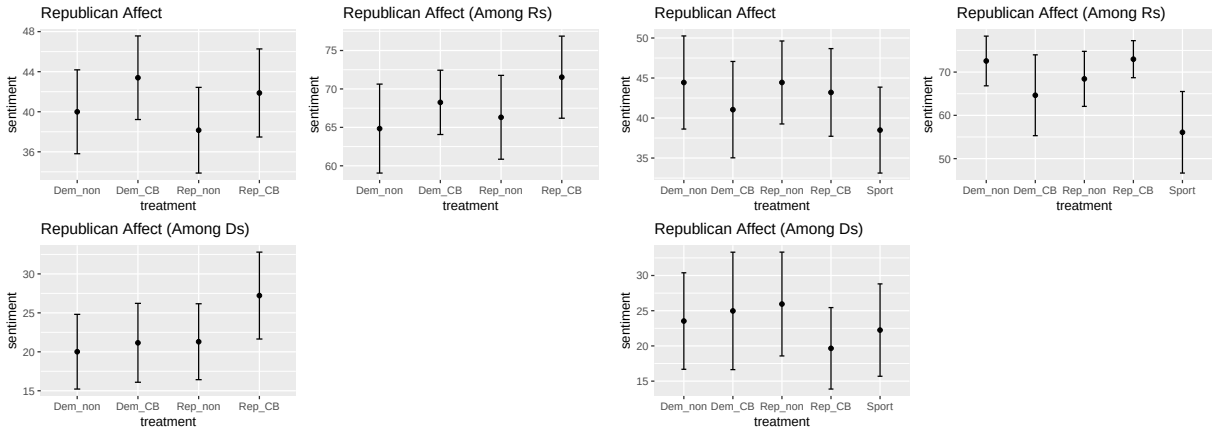
Further evidence that the treatments did not have the hypothesized effects and that this was not due to a lack of uptake comes from the factual knowledge questions. Figure 3.2 displays these results. There were three information retention questions in both Experiment 1 and Experiment 2; the latter added a sports-related placebo information retention question.[5]

Experiment 1, on top, again shows a lack of significant results. The largest effect size on the entire population, in fact, is in the *opposite* of the expected direction: the Democrat-leaning non-clickbait headline caused an *increase* in information retention. The results for Experiment 2 are the same, except that subjects in the placebo condition answered far fewer

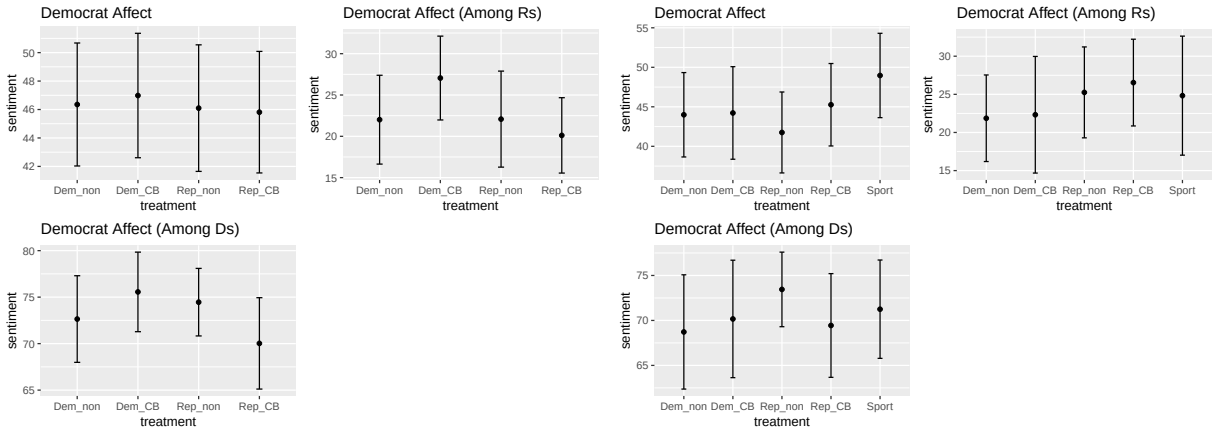
[5]All of the information retention questions were based on information provided in the body of the treatment news story for that Experiment. Because these were taken from existing news stories based on recent political news stories, it is possible that respondents could have known the correct answers *ex ante*. To minimize this problem, the questions concerned specific details from the stories that were not particularly salient to the overall political discussion.

Figure 3.1: Effects of Clickbait on Partisan Affect

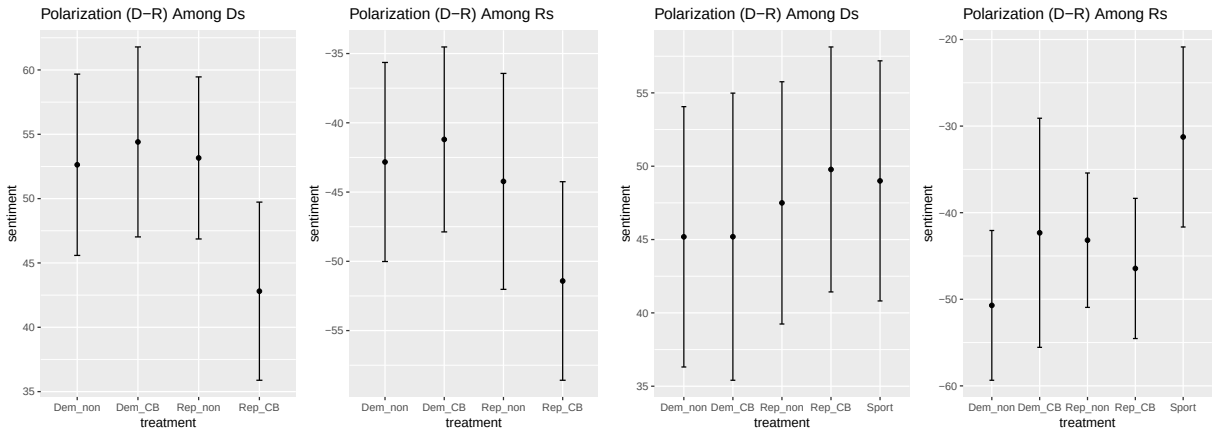
Experiment 1 **Republican Affect** Experiment 2



Experiment 1 **Democrat Affect** Experiment 2



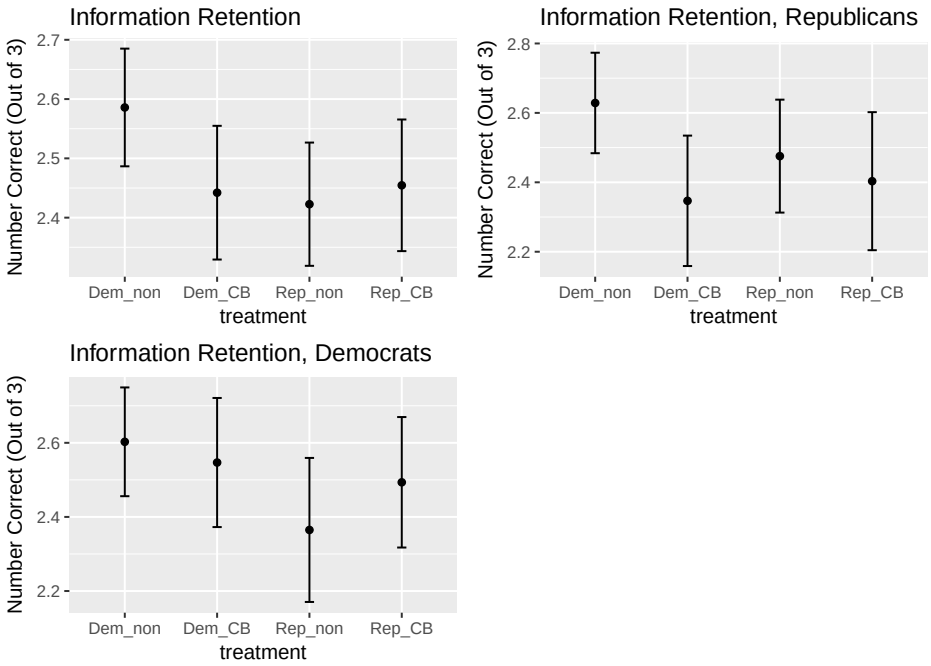
Experiment 1 **Affective Polarization** Experiment 2



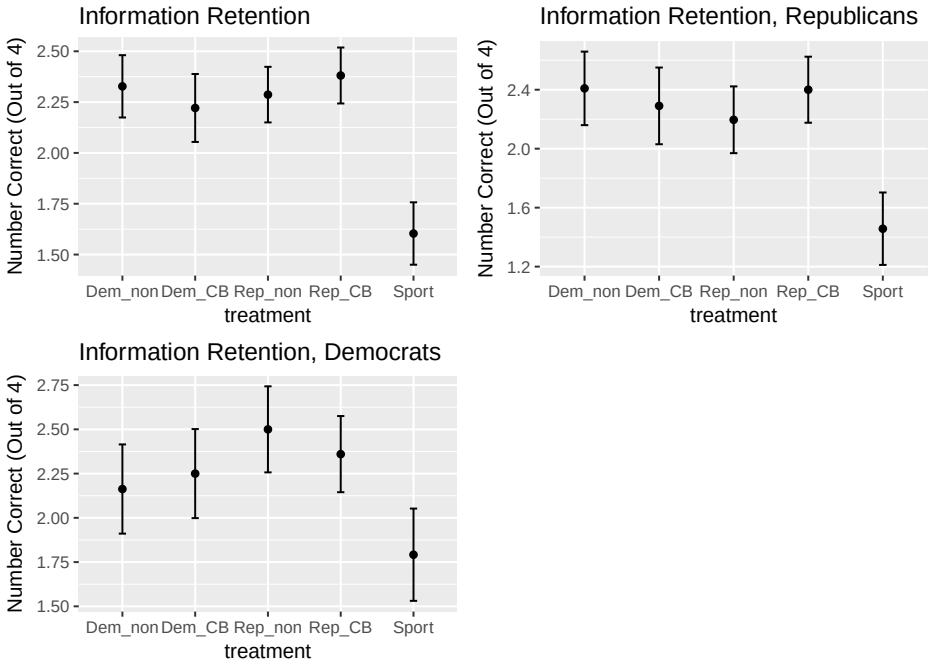
Bars represent 95% confidence intervals.

Figure 3.2: Effects of Clickbait on Information Retention

Experiment 1



Experiment 2



Bars represent 95% confidence intervals.

questions correctly. This is evidence that subjects were in fact reading the stories carefully and retaining the information, rather than relying on their *ex ante* knowledge.

The final hypothesized effect was on subjects' trust in both offline media and online media. These categories were explained to subjects in the questions:

```
In general, how much trust and confidence do you have in the offline mass
media -- such as newspapers, T.V. and radio -- when it comes to reporting
the news fully, accurately, and fairly -- a great deal, a fair amount, not
very much, or none at all?

In general, how much trust and confidence do you have in online-only media
--  such as blogs and online-only news websites -- when it comes to reporting
the news fully, accurately, and fairly -- a great deal, a fair amount, not
very much, or none at all?
```

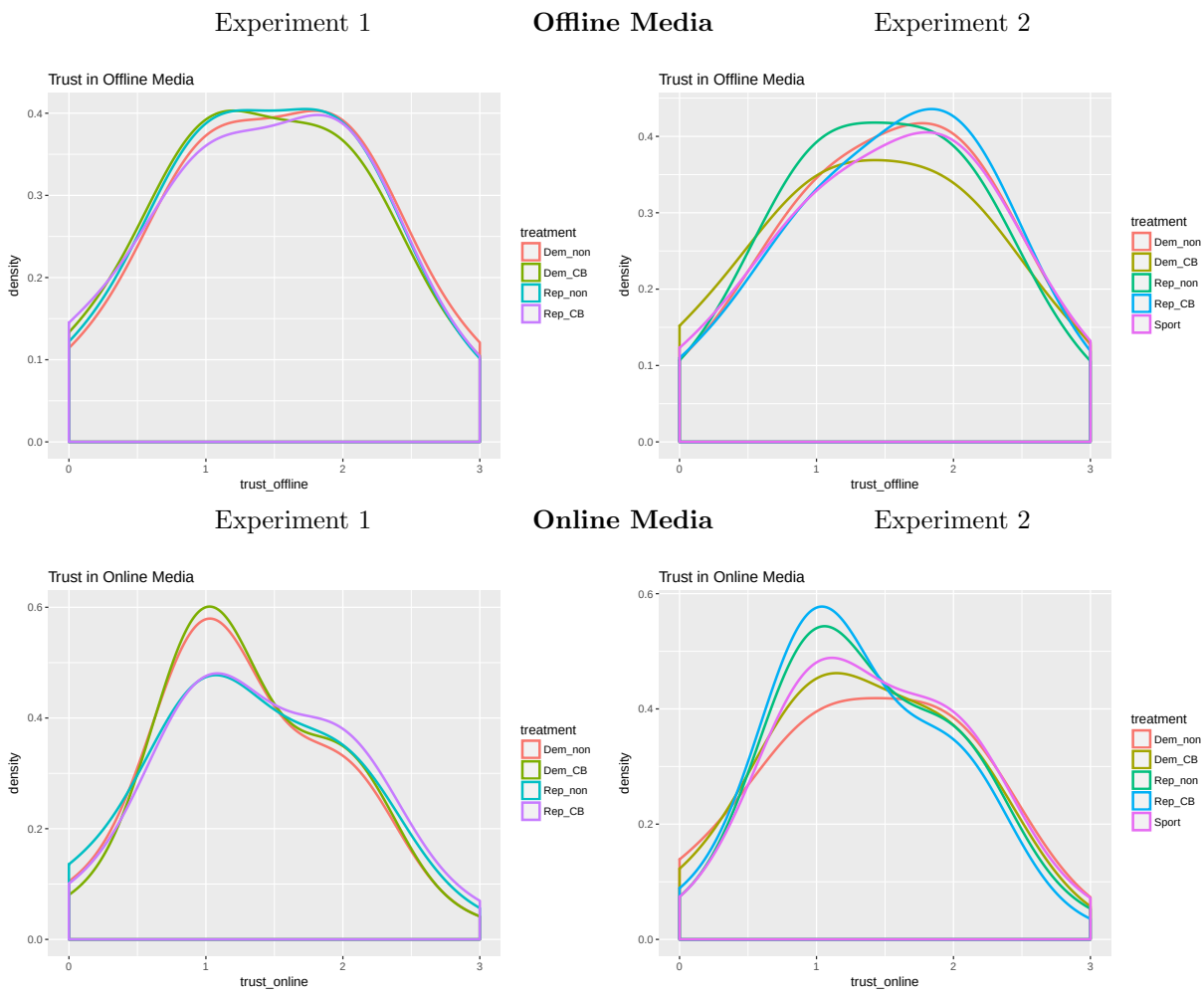
Figure 3.3 displays these results. Because this is a categorical four-point scale, the kernel plot is best able to show any treatment effects over the entire distribution of responses. There are no significant treatment effects. There are small spikes in the percentage of respondents rating their trust in Online Media (bottom panel) as "not very much" in both Experiments, but these differences are associated with the *opposite* treatment conditions: the two Democrat conditions in Experiment 1, and the two Republican conditions in Experiment 2.

In Experiment 2, subjects in the placebo condition evaluated both forms of media no differently than did subjects in the treatment conditions, in contrast to the results for information retention and Republican affect above. This is evidence that the placebo condition functioned as intentioned, and did not merely confuse subjects.[6]

These experimental results failed to provide any evidence in support of the hypothesized

[6]This was a concern because 6 of the 149 subjects in the placebo condition sent a message saying they had been surprised by the presence of a story about sports just before a battery of questions about the jobs report. Their main concern seems to have been that they not be penalized for giving incorrect answers.

Figure 3.3: Effects of Clickbait on Trust in Media



Kernel plots displaying subjects' evaluations of the trustworthiness of offline and online media. X-axis scale goes from "none at all" to "a great deal."

mechanisms driving affect/social polarization (point 5) or divergent information environments (point 6) in Figure 0.2 above. My primary explanation for these null results is that these hypothesized mechanisms (and the political trends they drive) are far more effective on a subset of the population of internet users. As I argue above, digital literacy is the crucial determinant of the heterogeneous effects of internet use.

For this reason, Mechanical Turk users may have been precisely the wrong population from which to sample; regardless of how representative they may be on other observable dimensions, all of these subjects necessarily possess sufficient technical savvy to complete the survey. Brewer, Morris and Piper (2016) provide explicit evidence that this is the case: they sample older adults who have not used Mechanical Turk and encourage them to perform example tasks on the platform, engaging in in-depth interviews throughout.

The vast majority of this sample of adults over 65 reported having used the internet for more than 15 years, reported being comfortable using computers, and placed themselves as "intermediate" on a three-point scale of familiarity with the Internet (p2250). However, the specific setup employed in the experiments described above would have posed a problem for many of them:

Tasks may ask the worker to take a survey and present an unlinked URL that the worker would have to copy and paste to a new tab or window to open. However, many participants were not familiar or comfortable with opening content in new tabs/windows, resulting in questions such as, "How do I get back to the instructions?" (P7) after a new tab was opened. Also, participants often forgot the instructions immediately upon opening the new window, particularly long and detailed instructions. P3 explained: "There's too many things to remember all at once. I mean, you don't–you can't read all these instructions and process it! I wouldn't remember them. I'd have to go back. I would never be able to remember all that... One of my complaints about some things on a computer is that, you know, if there's a bunch of instructions or stuff to know–and you have to open up a box and then if you go back to what you're working on the box is gone, and you can't just look up and reference it." (p2251)

Qualitative study of the way that different populations (especially along the dimension of age) engage with the internet and social media is increasingly essential. It is nearly impossible for a proficient internet user to appreciate the extent of the challenge posed by "opening content in new tabs/windows" for someone much less internet proficient. It is tempting to look to our own experiences to begin to study the experiences of others, but in the case of a technology as inherently heterogeneous as social media, this introspection will necessarily lead scholars astray.

3.4 Conclusion

The experimental results presented in this Chapter failed to provide evidence for the hypothesized mechanism driving affect polarization and differential information diets; however, these results failed to provide evidence for much of anything at all.

The main conclusion I have drawn from both the overview of the literature in the Introduction and these empirical results (as well as the results presented in Chapters 1 and 2) is that the behavioral/attitudinal impact of social media use is always and everywhere *heterogeneous*.

The variable heterogeneity of the effects of different media technologies is well established. The clearest example comes from Prior (2007), who conceptualizes two populations of television consumers: those with a high preference for entertainment (PfE), who will always chose to watch non-news programs, and those with low PfE. In the broadcast era, these groups were indistinguishable because of the lack of choice among the three broadcast providers. Broadcast television thus had a relatively *homogeneous* effect on viewers' political attitudes and information levels. With the advent of cable television, however, people with high PfE avoided news programs. The effect of cable television viewing was thus *heterogeneous* in the viewer's PfE; cable television led to a more polarized electorate as moderates became less politically engaged.

Changing the number of images simultaneously possible to view from 3 to 50 (broadcast to cable television) increased the heterogeneity of the effects of television. The internet and social media have made that number of possible images essentially infinite; you can never step in the same News Feed twice.

Heterogeneity should thus be central to any study of media or persuasive effects on social media. Average treatment effects on a representative population might well be expected to be zero.

There is growing evidence for this view in the context of political engagement. Using web-tracking data, Guess, Nyhan and Reifler (2017) conclude that the average number of times US internet users viewed Fake News during the 2016 election was quite low. This average masks the fact that "almost six in ten visits to Fake News websites came from the 10% of Americans with the most conservative information diets."

The dimension of heterogeneity most central to the theory I have developed in this paper—and which I intend to study further in the coming years—is in digital literacy. Scholars in Communications and Sociology have long studied what Eszther Hargittai has deemed the "Second-level Digital Divide" (Hargittai, 2001), and have developed survey instruments designed to estimate respondents' digital literacy Hargittai (2005); Van Deursen, Helsper and Eynon (2016). My aim is to incorporate these measures into mainstream Political Science research and make them essential to the study of the effects of the internet and social media.[7] Another implication of the focus on heterogeneity is the value in merely describing the contours of that heterogeneity, like the results presented in Table 3.2.

I conclude with a look at future cultural/technological trends that will (I believe) structure the intersection of social media and politics in the future. The most pressing development in this story is the widespread adoption of video as the primary medium for news content. Facebook and Twitter have both rolled out embedded videos that play automatically ("pre-roll"), and high-speed 3G and 4G internet service is now sufficiently widespread in the US

[7] I regret not having done so in the experimental analysis presented above. My ignorance of this literature was unfortunate, and gives me all the greater impulse to overcome existing disciplinary boundaries in this area.

that a majority of consumers can actually load these videos.

The tech news has begun reporting on this trend. "As Online Video Surges, Publishers Turn to Automation" ; "As Facebook Focuses on Video, Engagement for Top Publishers Declines"; "Mark Zuckerberg: Within Five Years, Facebook Will be Mostly Video".

Note especially "Facebook is predicting the end of the written word", in which a Facebook executive is quoted as saying that "In five years time Facebook 'will be definitely mobile, it will be probably all video.'"

Scholarly attention to this phenomenon is very new: although Pew has been studying trends in social media use for years, the most recent report was the first to ask about the use of Youtube (Smith and Anderson, 2018). Although Facebook use has plateaued over the past two years at 68% of adults, 73% report using Youtube, making it the most widely used social media site. Further, 94% of people aged 18-24 report using YouTube, making that the highest-penetration demographic of any in the sample.

Although YouTube is different from other social media sites in its structure, the essential features—the recommendation algorithm and the massive amount and breadth of available video content—are still in place. The "credibility cascade" I describe above thus applies to the economics of YouTube as well. The recommendation algorithm is less direct for YouTube than for Facebook, but the central feature that more popular videos are recommended to people YouTube thinks will want to watch them remains the same. The popularity of a given video or "channel" (what YouTube calls an entity which uploads videos) is often central to how that video or channel is discussed.

YouTube differs from Facebook in two other aspects that I find theoretically important. First, video is simply easier to consume than written news, causing a portion of the population that would not otherwise consume news to be exposed to it. Second, each video comes with a number of easily clickable "recommended videos" next to it, one of which will "autoplay" after a given video concludes. Technology writer James Bridle explored the horrifying consequences of these algorithms in the context of "Kids YouTube".

Video producers use figures popular among toddlers (Disney princesses and superheros) and put them in nonsensical, scatological or even violent scenarios. Toddlers, left alone with an iPad, click on the videos that look most appealing them, granting more views and more advertising revenue to more extreme videos. Video producers (and much of this process is entirely automated, using cheap computer graphics software) respond to this demand and create more extreme, nonsensical content. As Bridle describes one vivid example:

Familiar characters, nursery tropes, keyword salad, full automation, violence, and the very stuff of kids' worst dreams. And of course there are vast, vast numbers of these videos. Channel after channel after channel of similar content, churned out at the rate of hundreds of new videos every week. Industrialised nightmare production.

An analogous process seems to be going on for political content. YouTube (owned by Google) has seen far less attention among Political Scientists than Facebook or Twitter because a) YouTube plays a small role in how Political Scientists access political information and b) YouTube is extremely stingy with data. The most comprehensive study of the role of YouTube in the 2016 campaign (of which I'm aware) was conducted by an ex-Google employee for *The Guardian* newspaper.

Lewis and McCormick (2018) describe this experiment, in which an algorithm was "seeded" with an even number of pro-Trump and pro-Clinton videos and then selected one of the top five recommended videos "up next" iteratively. Each run was "fresh", without any search history.

The Guardian compiled a list of the most commonly recommended videos according to this process. The top 10 include "Must Watch!! Hillary Clinton tried to ban this video" and "TRUMP: the COMING LANDSLIDE Ancient Prophecy Documentary of Donald Trump / 2016." Overall, the most popular videos were actually those critical of Clinton; content analysis found that of the 643 videos coded as having bias, 86% favored Trump.

Why would this be? Clickbait economics tells us that in low-cost, low-credibility contexts like these, content popularity is key; the algorithm's aim is to give people content they

want to consume. YouTube responded to the publication of this study by saying the finding reflects "not a bias towards any particular candidate [but] a reflection of viewer interest."

The extreme partisan skew of low-credibility YouTube videos (relative to low-credibility written content) stems from the fact that *watching a video is easier than reading.*

The degree of digital literacy required to click a recommended video is zero; navigating Facebook is more difficult than navigating YouTube, and the former has higher social stakes (so that a more digitally literate niece might be able to help). But there's also the issue of straight-up *literacy*: for many people, reading is taxing and difficult in a way that watching a video is not!

The technology of video production and distribution has not yet become as cheap as written content. Interestingly, the bottleneck input here might be the technical capacity of content *producers*: video production is a less ubiquitous skill than text production. The widespread use of YouTube among the youngest generation ensures that this will not be the case forever.

Online persuasion will only continue to grow in importance as the technologies mature and new generations of internet-savvy individuals displace the least savvy older generations. Although careful empirical studies are essential to the understanding of this persuasion for political outcomes, each individual study also runs the risk of becoming rapidly outdated. A structural understanding of the incentives driving the production and distribution of online content is thus also important in enabling scholars to design empirical studies that are best able to inform our knowledge of enduring mechanisms of online persuasion.

Appendix for Chapter 3

C.1: Survey Instrument

Suite 804, New York, New York, 10012, at ask.humansubjects@nyu.edu or (212) 998-4808.

You may only participate in this study once; multiple responses from the same Worker ID will be discarded and you will not be compensated.

Do you agree to participate in this study?

- ○ Yes

Block 1

Background information

Before we get started, we'd like to hear a little bit about your background.

How old are you?

What is your gender?

- ○ Male
- ○ Female
- ○ Other

What is the highest level of education you've completed?

- ○ Have not finished high school
- ○ High school
- ○ Some college
- ○ College
- ○ Postgraduate degree

How often do you use the Internet?

- Pretty much all the time
- Several times a day
- About once a day
- 3 to 6 days a week
- 1 to 2 days a week
- Every few weeks
- Less often

How often do you use Facebook?

- Pretty much all the time
- Several times a day
- About once a day
- 3 to 6 days a week
- 1 to 2 days a week
- Every few weeks
- Less often
- I don't use Facebook

How often do you use Twitter?

- Pretty much all the time
- Several times a day
- About once a day
- 3 to 6 days a week
- 1 to 2 days a week
- Every few weeks
- Less often
- I don't use Twitter

Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent, or what?

- ○ Democrat
- ○ Lean Democrat
- ○ Independent
- ○ Lean Republican
- ○ Republican

How often do you read news stories online?

- ○ Pretty much all the time
- ○ Several times a day
- ○ About once a day
- ○ 3 to 6 days a week
- ○ 1 to 2 days a week
- ○ Every few weeks
- ○ Less often

How often do you read news stories offline (in the newspaper, printed news magazines)?

- ○ Pretty much all the time
- ○ Several times a day
- ○ About once a day
- ○ 3 to 6 days a week
- ○ 1 to 2 days a week
- ○ Every few weeks
- ○ Less often

Media Choice

In this section, we're going to present you with several hypothetical headlines. If you could only choose to read one of these stories, which would it be?

- ○ Warriors, Steph Curry agree to 5-year, $201M deal that's richest in NBA history
- ○ People are loving this: President Trump dismisses mainstream media frenzy about Russia allegations
- ○ Report shows that climate change is much worse than previously feared
- ○ 'The Grand Tour' host Richard Hammond injured in car crash

These page timer metrics will not be displayed to the recipient.
First Click: *0 seconds*
Last Click: *0 seconds*
Page Submit: *0 seconds*
Click Count: *0 clicks*

If you could only choose to read one of these stories, which would it be?

- ○ Sorry? Justin Bieber is straight up banned from performing in China now
- ○ Planned Parenthood on the ropes as funding slashed
- ○ Leaderboard: Spieth battling weather at The Open
- ○ This is why President Trump's plan for dealing with North Korea is a disaster

These page timer metrics will not be displayed to the recipient.
First Click: *0 seconds*
Last Click: *0 seconds*
Page Submit: *0 seconds*
Click Count: *0 clicks*

These page timer metrics will not be displayed to the recipient.
First Click: *0 seconds*
Last Click: *0 seconds*
Page Submit: *0 seconds*
Click Count: *0 clicks*

If you could only choose to read one of these stories, which would it be?

- Media investigations keep finding new information about Russia allegations
- Gurriel, Astros rally to get past Yankees
- Pop star Ariana Grande visits fans in hospital
- This is what the Mexican border wall will empower local governments to do

If you could only choose to read one of these stories, which would it be?

- Louisville, North Carolina and the future of NCAA punishing power
- This will make you furious: new email evidence shows Donald Trump, Jr., conspired with the Russians
- Brad Pitt is all smiles, looks buff at his art studio
- Supreme Court allows the popular travel ban to take effect

These page timer metrics will not be displayed to the recipient.

First Click: *0 seconds*

Last Click: *0 seconds*

Page Submit: *0 seconds*

Click Count: *0 clicks*

If you could only choose to read one of these stories, which would it be?

- Chester Bennington, Linkin Park frontman, dead at 41
- Mexican border wall to nearly bankrupt local government
- After 8 years, OJ Simpson released on parole
- Democrats are freaking out: evidence from Seattle shows raising minimum wages causes unemployment

These page timer metrics will not be displayed to the recipient.

First Click: *0 seconds*

Last Click: *0 seconds*

Page Submit: *0 seconds*

Click Count: *0 clicks*

If you could only choose to read one of these stories, which would it be?

- CNN tweets new response to controversy
- Here's what happened with the Dallas Cowboys this weekend
- Drake and Rihanna are getting back together after a vacation in Area 51
- Survey taker: always select this option, ignore the other three headlines

These page timer metrics will not be displayed to the recipient.
First Click: *0 seconds*
Last Click: *0 seconds*
Page Submit: *0 seconds*
Click Count: *0 clicks*

If you could only choose to read one of these stories, which would it be?

- 5 things you need to know about the Supreme Court's cowardly travel ban approval
- Donald Trump, Jr., under attack for doing nothing wrong
- Pats owner Robert Kraft says NFL's future is through over the top
- Snoop Dogg admits he downloaded a 'Bootleg' of Jay-Z's new album '4:44'

These page timer metrics will not be displayed to the recipient.
First Click: *0 seconds*
Last Click: *0 seconds*
Page Submit: *0 seconds*
Click Count: *0 clicks*

If you could only choose to read one of these stories, which would it be?

- Blac Chyna's attorney says he's 'Exploring All Legal Remedies' after her ex Rob Kardashian posted explicit photos
- President Trump's bold new plan for dealing with North Korea
- Sports world shocked by decision in Manny Pacquiao-Jeff Horn fight
- Republicans are shocked to see that slashing funding to Planned Parenthood increases abortion rate

These page timer metrics will not be displayed to the recipient.
First Click: *0 seconds*

Last Click: *0 seconds*
Page Submit: *0 seconds*
Click Count: *0 clicks*

If you could only choose to read one of these stories, which would it be?

- ○ Seattle's minimum wage increase shows that paying people more actually works
- ○ Jay-Z's new album admits that he cheated on Beyoncé
- ○ 10 reasons that climate change isn't a serious problem
- ○ Germany bests Chile in warmup to next year's World Cup

These page timer metrics will not be displayed to the recipient.
First Click: *0 seconds*
Last Click: *0 seconds*
Page Submit: *0 seconds*
Click Count: *0 clicks*

Block 4

In this section, we're going to ask you to read a news story, paying careful attention to detail. Please click the following link and read the news story it links to. When you're done reading, close that tab and continue this survey.

Trump economic policies not working

- ○ I've read the story carefully, and I'm ready to continue the survey

In this section, we're going to ask you to read a news story, paying careful attention to detail. Please click the following link and read the news story it links to. When you're done reading, close that tab and continue this survey.

Trump economic policies working

- ○ I've read the story carefully, and I'm ready to continue the survey

In this section, we're going to ask you to read a news story, paying careful attention to detail. Please click the following link and read the news story it links to. When you're done reading, close that tab and continue this survey.

Democrats won't like this economic news: Trump policies working!

○ I've read the story carefully, and I'm ready to continue the survey

In this section, we're going to ask you to read a news story, paying careful attention to detail. Please click the following link and read the news story it links to. When you're done reading, close that tab and continue this survey.

Republicans won't like this economic news: Trump policies not working!

○ I've read the story carefully, and I'm ready to continue the survey

In this section, we're going to ask you to read a news story, paying careful attention to detail. Please click the following link and read the news story it links to. When you're done reading, close that tab and continue this survey.

See who LeBron James and Steph Curry picked for their All-Star teams

○ I've read the story carefully, and I'm ready to continue the survey

A little more background information

In general, how much trust and confidence do you have in the offline mass media -- such as newspapers, T.V. and radio -- when it comes to reporting the news fully, accurately, and fairly -- a great deal, a fair amount, not very much, or none at all?

○ A great deal

- ○ A fair amount
- ○ Not very much
- ○ None at all

In general, how much trust and confidence do you have in online-only media -- such as blogs and online-only news websites -- when it comes to reporting the news fully, accurately, and fairly -- a great deal, a fair amount, not very much, or none at all?

- ○ A great deal
- ○ A fair amount
- ○ Not very much
- ○ None at all

We'd like to get your feelings toward certain groups related to US politics. Ratings between 50 and 100 degrees mean that you feel favorable and warm toward the group. Ratings between 0 and 50 degrees mean that you don't feel favorable toward the group and that you don't care too much for that group . You would rate the group at 50 degrees if you don't feel particularly warm or cold toward the group.

	0	10	20	30	40	50	60	70	80	90	100
Democrats											
Republicans											
Online-only media											
Traditional media											

These page timer metrics will not be displayed to the recipient.
First Click: *0 seconds*
Last Click: *0 seconds*
Page Submit: *0 seconds*
Click Count: *0 clicks*

Questions about the Republican health care bill

The October economy report indicated that the unemployment rate:

- Is lower than it was a year ago
- Is pretty much the same as it was a year ago
- Is higher than it was a year ago

The October economy report indicated that the number of people either employed or actively looking for work:

- fell
- stayed pretty much the same
- increased

The October economy report indicated that wages for non-managers:

- fell
- stayed pretty much the same
- increased

Which team is Kevin Durant on for the NBA All-Star game?

- Team Steph Curry
- Team LeBron James
- Team Michael Jordan

Bibliography

Abramowitz, Alan I and Kyle L Saunders. 2006. "Exploring the bases of partisanship in the American electorate: Social identity vs. ideology." *Political Research Quarterly* 59(2):175–187.

Abramowitz, Alan I and Steven Webster. 2016. "The rise of negative partisanship and the nationalization of US elections in the 21st century." *Electoral Studies* 41:12–22.

Acemoglu, Daron and James A Robinson. 2005. *Economic origins of dictatorship and democracy*. Cambridge University Press.

Achen, Christopher H and Larry M Bartels. 2016. *Democracy for realists: Why elections do not produce responsive government*. Princeton University Press.

Aday, Sean et al. 2010. *Blogs and bullets: New media in contentious politics*. United States Institute of Peace.

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa and Ekaterina Zhuravskaya. 2013. "Radio and the rise of Nazis in pre-war Germany." *Available at SSRN 2242446* .

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa and Ekaterina Zhuravskaya. 2014. "Radio and the Rise of the Nazis in Prewar Germany." *Available at SSRN 2242446* .

Agarwal, Sheetal D, W Lance Bennett, Courtney N Johnson and Shawn Walker. 2014. "A model of crowd enabled organization: Theory and methods for understanding the role of twitter in the occupy protests." *International Journal of Communication* 8:27.

Aldrich, John H. 1993. "Rational choice and turnout." *American Journal of Political Science* pp. 246–278.

Allcott, Hunt and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Technical report National Bureau of Economic Research.

Allport, Gordon Willard. 1954. *The Nature of Prejudice*. Basic Books.

Alonzo, Mei and Milam Aiken. 2004. "Flaming in electronic communication." *Decision Support Systems* 36(3):205–213.

Andersen, Robert, James Tilley and Anthony F Heath. 2005. "Political knowledge and enlightened preferences: party choice through the electoral cycle." *British Journal of Political Science* 35(02):285–302.

Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg and Jure Leskovec. 2012. Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM pp. 703–712.

Angeletos, George-Marios, Christian Hellwig and Alessandro Pavan. 2006. "Signaling in a global game: Coordination and policy traps." *Journal of Political Economy* 114(3):452–484.

Aral, Sinan and Dylan Walker. 2012. "Identifying influential and susceptible members of social networks." *Science* 337(6092):337–341.

Arceneaux, Kevin and Martin Johnson. 2013. *Changing minds or changing channels?: Partisan news in an age of choice*. University of Chicago Press.

Bächtiger, André, Simon Niemeyer, Michael Neblo, Marco R Steenbergen and Jürg Steiner. 2010. "Disentangling diversity in deliberative democracy: Competing theories, their blind spots and complementarities." *Journal of Political Philosophy* 18(1):32–63.

Bakker, Ryan, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen and Milada Vachudova. 2015. "2014 Chapel Hill Expert Survey." 2015.1.

Bakshy, Eytan, Solomon Messing and Lada A Adamic. 2015. "Exposure to ideologically diverse news and opinion on Facebook." *Science* 348(6239):1130–1132.

Banks, Antoine J. 2014. "The public's anger: White racial attitudes and opinions toward health care reform." *Political Behavior* 36(3):493–514.

Banks, Antoine J. 2016. "Are Group Cues Necessary? How Anger Makes Ethnocentrism Among Whites a Stronger Predictor of Racial and Immigration Policy Opinions." *Political Behavior* pp. 1–23.

Barabas, Jason and Jennifer Jerit. 2009. "Estimating the Causal Effects of Media Coverage on Policy-Specific Knowledge." *American Journal of Political Science* 53(1):73–89.

Barabas, Jason, Jennifer Jerit, William Pollock and Carlisle Rainey. 2014. "The question (s) of political knowledge." *American Political Science Review* 108(04):840–855.

Barberá, Pablo. 2013. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Proceedings of the Social Media and Political Participation, Florence, Italy* pp. 10–11.

Barberá, Pablo. 2014. "How social media reduces mass political polarization. Evidence from Germany, Spain, and the US." *Job Market Paper, New York University* .

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76–91.

Barberá, Pablo and Gonzalo Rivero. 2014. "Understanding the political representativeness of Twitter users." *Social Science Computer Review* p. 0894439314558836.

Barberá, Pablo, John T Jost, Jonathan Nagler, Joshua A Tucker and Richard Bonneau. 2015. "Tweeting From Left to Right Is Online Political Communication More Than an Echo Chamber?" *Psychological science* p. 0956797615594620.

Barberá, Pablo, Ning Wang, Richard Bonneau, John T Jost, Jonathan Nagler, Joshua Tucker and Sandra González-Bailón. 2015. "The critical periphery in the growth of social protests." *PloS one* 10(11):e0143611.

Barbera, Pablo, Richard Bonneau, John T Jost, Jonathan Nagler and Joshua Tucker. 2013. "Is There Anybody Out There? The Effects of Legislators' Communication with their Constituents.".

Barberá, Pablo, Richard Bonneau, Patrick Egan, John T Jost, Jonathan Nagler and Joshua Tucker. 2014. Leaders or followers? Measuring political responsiveness in the US Congress using social media data. In *110th American Political Science Association Annual Meeting*.

Baron, David P. 2006. "Persistent media bias." *Journal of Public Economics* 90(1):1–36.

Bartels, Larry M. 1988. *Presidential primaries and the dynamics of public choice*. Princeton University Press.

Bartels, Larry M. 1996. "Uninformed votes: Information effects in presidential elections." *American Journal of Political Science* pp. 194–230.

Baum, Matthew A and Samuel Kernell. 1999. "Has cable ended the golden age of presidential television?" *American Political Science Review* 93(01):99–114.

Baumgartner, Frank R. 2001. "Political agendas.".

Baumgartner, Frank R and Bryan D Jones. 2010. *Agendas and instability in American politics*. University of Chicago Press.

Baumgartner, Frank R, Bryan D Jones and Peter B Mortensen. 2014. "Punctuated equilibrium theory: Explaining stability and change in public policymaking." *Theories of the policy process* pp. 59–103.

Baumol, William J. 1986. "Contestable markets: an uprising in the theory of industry structure." *Microtheory: applications and origins* pp. 40–54.

Baumol, William J, John C Panzar and Robert D Willig. 1983. "Contestable markets: An uprising in the theory of industry structure: Reply." *The American Economic Review* 73(3):491–496.

Baumol, William J, John C Panzar, Robert D Willig, Elizabeth E Bailey, Dietrich Fischer and Dietrich Fischer. 1982. "Contestable markets and the theory of industry structure.".

BBC. 2014. "Twitter confirma bloqueo de imagenes en Venezuela." *bbc.com.uk* .

Beam, Randal A, David H Weaver and Bonnie J Brownlee. 2009. "Changes in professionalism of US journalists in the turbulent twenty-first century." *Journalism & Mass Communication Quarterly* 86(2):277–298.

Beauchamp, Nick. 2013. Predicting and interpolating state-level polling using twitter textual data. In *Meeting on automated text analysis, London School of Economics, London*.

Bejan, Teresa M. 2017. *Mere Civility*. Harvard University Press.

Bennett, W Lance, Alexandra Segerberg and Shawn Walker. 2014. "Organization in the crowd: peer production in large-scale networked protests." *Information, Communication & Society* 17(2):232–260.

Berry, Jeffrey M and Sarah Sobieraj. 2013. *The outrage industry: Political opinion media and the new incivility*. Oxford University Press.

Bertot, John C, Paul T Jaeger and Justin M Grimes. 2010. "Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies." *Government Information Quarterly* 27(3):264–271.

Bertrand, Marianne and Sendhil Mullainathan. 2003. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *National*

Bureau of Economic Research .

Bimber, Bruce, Andrew Flanagin and Cynthia Stohl. 2012. *Collective action in organizations: Interaction and engagement in an era of technological change*. Cambridge University Press.

Binder, Jens, Hanna Zagefka, Rupert Brown, Friedrich Funke, Thomas Kessler, Amelie Mummendey, Annemie Maquil, Stephanie Demoulin and Jacques-Philippe Leyens. 2009. "Does contact reduce prejudice or does prejudice reduce contact? A longitudinal test of the contact hypothesis among majority and minority groups in three European countries." *Journal of Personality and Social Psychology* 96(4):843.

Bishop, Jonathan. 2013. "The effect of de-individuation of the Internet Troller on Criminal Procedure implementation: An interview with a Hater." *International Journal of Cyber Criminology* 7(1).

Blanchard, Fletcher A, Christian S Crandall, John C Brigham and Leigh Ann Vaughn. 1994. "Condemning and condoning racism: A social context approach to interracial settings." *Journal of Applied Psychology* 79(6):993.

Blau, David M and Bruce A Weinberg. 2017. "Why the US science and engineering workforce is aging rapidly." *Proceedings of the National Academy of Sciences* p. 201611748.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *the Journal of machine Learning research* 3:993–1022.

Boas, Taylor C and Shanti Kalathil. 2003. "Open networks, closed regimes: The impact of the Internet on authoritarian rule." *Washington, DC: Carnegie Endowment* .

Bode, Leticia and Emily K. Vraga. 2015. "In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media." *Journal of Communication* 65(4):619–638.
URL: *http://dx.doi.org/10.1111/jcom.12166*

Bohdanova, Tetyana. 2014. "Unexpected revolution: the role of social media in Ukraine's Euromaidan uprising." *European View* pp. 1–10.

Boix, Carles and Milan W Svolik. 2013. "The foundations of limited authoritarian government: Institutions, commitment, and power-sharing in dictatorships." *The Journal of Politics* 75(02):300–316.

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle and James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature* 489(7415):295.

Bond, Robert M, Jaime E Settle, Christopher J Fariss, Jason J Jones and James H Fowler. 2017. "Social endorsement cues and political participation." *Political Communication* 34(2):261–281.

Bordia, Prashant. 1997. "Face-to-face versus computer-mediated communication: A synthesis of the experimental literature." *Journal of Business Communication* 34(1):99–118.

Boxell, Levi, Matthew Gentzkow and Jesse M Shapiro. 2017. Is the internet causing political polarization? Evidence from demographics. Technical report National Bureau of Economic Research.

Boydstun, Amber E. 2013. *Making the news: Politics, the media, and agenda setting*. University of Chicago Press.

Brader, Ted. 2005. "Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions." *American Journal of Political Science* 49(2):388–405.

Bretschneider, Uwe, Thomas Wöhner and Ralf Peters. 2014. "Detecting Online Harassment in Social Networks.".

Brewer, Marilynn B. 1999. "The psychology of prejudice: Ingroup love and outgroup hate?" *Journal of Social Issues* 55(3):429–444.

Brewer, Robin, Meredith Ringel Morris and Anne Marie Piper. 2016. Why would anybody do this?: Understanding older adults' motivations and challenges in crowd work. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM pp. 2246–2257.

Buckels, Erin E, Paul D Trapnell and Delroy L Paulhus. 2014. "Trolls just want to have fun." *Personality and individual Differences* 67:97–102.

Bueno de Mesquita, Bruce, Alastair Smith, Randolph Siverson and James D. Morrow. 2005. *The logic of political survival*. MIT press.

Bullock, John G. 2011. "Elite influence on public opinion in an informed electorate." *American Political Science Review* 105(03):496–515.

Burden, Barry C and D Sunshine Hillygus. 2009. "Polls and Elections: Opinion Formation, Polarization, and Presidential Reelection." *Presidential Studies Quarterly* 39(3):619–635.

Bureau of Labor Statistics. 2016. *Employment trends in newspaper publishing and other media, 1990–2016*. BLS.

Burger, John D, John Henderson, George Kim and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics pp. 1301–1309.

Campbell, Angus, Philip E Converse, Warren E Miller and E Donald. 1960. "The American Voter.".

Campbell, Angus, Philip E Converse and Warren Miller. 1960. "The American Voter.".

Carpini, Michael X Delli and Scott Keeter. 1991. "Stability and Change in the US Public's Knowledge of Politics." *Public Opinion Quarterly* 55(4):583–612.

Carpini, Michael X Delli and Scott Keeter. 1993. "Measuring political knowledge: Putting first things first." *American Journal of Political Science* pp. 1179–1206.

Carpini, Michael X Delli and Scott Keeter. 1997. *What Americans know about politics and why it matters*. Yale University Press.

Carpini, Michael X Delli and Scott Keeter. 2000. "Gender and political knowledge." *Gender and American politics: Women, men, and the political process* pp. 21–52.

Cellan-Jones, Rory. 2017. *Facebook's News Feed experiment panics publishers*. BBC.com.

Chadwick, Andrew. 2006. *Internet politics: States, citizens, and new communication technologies*. Oxford University Press, USA.

Chan, Elizabeth. 2016. "Donald Trump, Pepe the frog, and white supremacists: an explainer.".

Chang, Linchiat and Jon A Krosnick. 2009. "National surveys via RDD telephone interviewing versus the internet comparing sample representativeness and response quality." *Public Opinion Quarterly* 73(4):641–678.

Chen, Adrian. 2015. "The Agency." *New York Times Magazine* June 2, 2015.

Chen, Jidong, Jennifer Pan and Yiqing Xu. 2015. "Sources of authoritarian responsiveness: A field experiment in china." *American Journal of Political Science* .

Chen, Vivian Hsueh Hua and Yuehua Wu. 2015. "Group identification as a mediator of the effect of players' anonymity on cheating in online games." *Behaviour & Information Technology* 34(7):658–667.

Chen, Ying, Yilu Zhou, Sencun Zhu and Heng Xu. 2012. "Detecting offensive language

in social media to protect adolescent online safety". In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE pp. 71–80.

Cheng, Justin, Cristian Danescu-Niculescu-Mizil and Jure Leskovec. 2015. "Antisocial Behavior in Online Discussion Communities." *arXiv preprint arXiv:1504.00680* .

Cheng, Justin, Michael Bernstein, Cristian Danescu-Niculescu-Mizil and Jure Leskovec. 2017. "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions.".

Cheng, Xueqi, Yanyan Lan, Jiafeng Guo and Xiaohui Yan. 2014. "BTM: Topic Modeling over Short Texts." *IEEE Transactions on Knowledge and Data Engineering* p. 1.

Chhibber, Pradeep and Jasjeet S Sekhon. 2014. "The asymmetric role of religious appeals in India.".

Christensen, Darin and Francisco Garfias. 2015. "Can You Hear Me Now?: How Communication Technology Affects Protest and Repression." *SSRN* .

Ciccariello-Maher, Georeg. 2014. "LaSalida? Venezuela at a Crossroads." *The Nation* .

Cobb, Roger W and Marc Howard Ross. 1997. *Cultural strategies of agenda denial: Avoidance, attack, and redefinition*. Univ Pr of Kansas.

Cobb, Roger William. 1983. *Participation in American politics: The dynamics of agenda-building*. Johns Hopkins University Press.

Cogburn, Derrick L and Fatima K Espinoza-Vasquez. 2011. "From networked nominee to networked nation: Examining the impact of Web 2.0 and social media on political participation and civic engagement in the 2008 Obama campaign." *Journal of Political Marketing* 10(1-2):189–213.

Cohen, Bernard C. 1963. "The press, the public and foreign policy." *Reader In Public Opinion and Communication* pp. 134–35.

Cohen, Marty, David Karol, Hans Noel and John Zaller. 2009. *The party decides: Presidential nominations before and after reform*. University of Chicago Press.

Coleman, Gabriella. 2014. *Hacker, hoaxer, whistleblower, spy: The many faces of Anonymous*. Verso Books.

Coleman, LH, CE Paternite and RC Sherman. 1999. "A reexamination of deindividuation in synchronous computer-mediated communication." *Computers in Human Behavior*

15(1):51–65.

Confessore, Nicholas and Karen Yourish. 2016. "2 Billion Worth of Free Media for Donald Trump." *Pew Research Center* 25.

Conover, Michael D, Bruno Gonçalves, Alessandro Flammini and Filippo Menczer. 2012. "Partisan asymmetries in online political activity." *EPJ Data Science* 1(1):1–19.

Converse, Philip E. 1969. "Of time and partisan stability." *Comparative political studies* 2(2):139.

Converse, Phillip. 1964. "The Nature of Belief Systems in Mass Publics. In Ideology and Discontent, ed. David Apter. New York: Free Press.".

Coppock, Alexander, Andrew Guess and John Ternovski. 2015. "When Treatments are Tweets: A Network Mobilization Experiment over Twitter." *Political Behavior* pp. 1–24.

Cormack, Lindsey. 2013. "Gender and Vote Revelation Strategy in Congress." *Unpublished Manuscript* .

Corrales, Javier. 2013. "Chavismo After Chavez." *Foreign Affairs* .

Corrales, Javier and Michael Penfold-Becerra. 2011. *Dragon in the tropics: Hugo Chavez and the political economy of revolution in Venezuela*. Brookings Institution Press.

Cox, Gary W. 2009. "Authoritarian elections and leadership succession." *Unpublished manuscript* .

Crandall, Christian S, Amy Eshleman and Laurie O'Brien. 2002. "Social norms and the expression and suppression of prejudice: the struggle for internalization." *Journal of personality and social psychology* 82(3):359.

Cranmer, Skyler J, Christopher T Dawes et al. 2012. "The heritability of foreign policy preferences." *Twin Research and Human Genetics* 15(1):52.

Curran, James, Shanto Iyengar, Anker Brink Lund and Inka Salovaara-Moring. 2009. "Media System, Public Knowledge and Democracy A Comparative Study." *European Journal of Communication* 24(1):5–26.

Dahl, Robert A. 2013. *A preface to democratic theory*. University of Chicago Press.

DÍAZ, SARA CAROLINA. 2014. "Sector de la oposición convoca a marcha para el 12 de febrero." *El Universal* .

de Mesquita, Ethan Bueno. 2010. "Regime change and revolutionary entrepreneurs." *American Political Science Review* 104(03):446–466.

Deibert, Ronald, John Palfrey, Rafal Rohozinski, Jonathan Zittrain and Miklos Haraszti. 2010. *Access controlled: The shaping of power, rights, and rule in cyberspace*. Mit Press.

DellaVigna, Stefano and Ethan Kaplan. 2006. The Fox News effect: Media bias and voting. Technical report National Bureau of Economic Research.

Department of Consumer Affairs, Board of Barbering and Cosmetology. 2016. "Frequently Asked Questions." *CA Website* .

DiMaggio, Paul, Eszter Hargittai et al. 2001. "From the 'digital divide'to 'digital inequality': Studying Internet use as penetration increases." *Princeton: Center for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University* 4(1):4–2.

DiMaggio, Paul and Walter W Powell. 1983. "The iron cage revisited: Collective rationality and institutional isomorphism in organizational fields." *American Sociological Review* 48(2):147–160.

Dinakar, Karthik, Roi Reichart and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. In *The Social Mobile Web*.

Dovidio, John F and Samuel L Gaertner. 1999. "Reducing prejudice combating intergroup biases." *Current Directions in Psychological Science* 8(4):101–105.

Downs, Anthony. 1957. "An economic theory of political action in a democracy." *The journal of political economy* pp. 135–150.

Duggan, M and A Smith. 2016. "The political environment on social media." *Pew Research Center* 25.

Duggan, Maeve. 2015. *The Demographics of Social Media Users*. Pew.

Earl, Jennifer, Heather McKee Hurwitz, Analicia Mejia Mesinas, Margaret Tolan and Ashley Arlotti. 2013. "This protest will be tweeted: Twitter and protest policing during the Pittsburgh G20." *Information, Communication & Society* 16(4):459–478.

Editors, The. 2015. "A Growth Spurt." *http://unabridged.merriam-webster.com/blog/2015/05/a-growth-spurt/* .

Edmond, Chris. 2013. "Information manipulation, coordination, and regime change." *The Review of Economic Studies* 80(4):1422–1458.

Esfandiari, Golnaz. 2010. "The Twitter Devolution." *Foreign Policy* 7:2010.

Evans, Geoffrey and Jon Mellon. 2015. "Working Class Votes and Conservative Losses: Solving the UKIP Puzzle." *Parliamentary Affairs* p. gsv005.

Facebook. 2016. "Choose your audience." *Facebook Advertising Website* .

Fieldhouse, E., J. Green, G. Evans, H. Schmitt and C. van der Eijk. 2015. "British Election Study Internet Panel Wave 6.".

Fischer, Peter, Eva Jonas, Dieter Frey and Stefan Schulz-Hardt. 2005. "Selective exposure to information: The impact of information limits." *European Journal of Social Psychology* 35(4):469–492.

Fischer, Peter, Stefan Schulz-Hardt and Dieter Frey. 2008. "Selective exposure and information quantity: how different information quantities moderate decision makers' preference for consistent and inconsistent information." *Journal of personality and social psychology* 94(2):231.

Fishkin, James S. 2011. *When the people speak: Deliberative democracy and public consultation*. Oxford University Press.

Flaxman, Seth, Sharad Goel and Justin M Rao. 2013. "Ideological segregation and the effects of social media on news consumption." *Available at SSRN 2363701* .

Flaxman, Seth, Sharad Goel and Justin M Rao. 2016. "Filter bubbles, echo chambers, and online news consumption." *Public Opinion Quarterly* 80(S1):298–320.

Fong, Christian and Justin Grimmer. 2016. "Discovery of Treatments from Text Corpora.".

Fowler, Anthony and Michele Margolis. 2014. "The political consequences of uninformed voters." *Electoral Studies* 34:100–110.

Francia, Peter L. 2017. "Free media and Twitter in the 2016 presidential election: The unconventional campaign of Donald Trump." *Social Science Computer Review* p. 0894439317730302.

Franklin, Charles H and John E Jackson. 1983. "The dynamics of party identification." *American Political Science Review* 77(04):957–973.

Friedman, Uri. 2014. "Why Venezuela's Revolution Will Be Tweeted." *The Atlantic* .

Friggeri, Adrien, Lada A Adamic, Dean Eckles and Justin Cheng. 2014. Rumor Cascades.

In *ICWSM*.

Frijda, Nico H. 1988. "The laws of emotion." *American psychologist* 43(5):349.

Gabler, Neil. 2016. The internet and social media are increasingly divisive and undermining of democracy. Technical report Alternet.org.

Gainous, Jason and Kevin M Wagner. 2014. *Tweeting to Power: The Social Media Revolution in American Politics*. Oxford University Press.

Gandhi, Jennifer and Ellen Lust-Okar. 2009. "Elections under authoritarianism." *Annual Review of Political Science* 12:403–422.

Garrett, R Kelly and Brian E Weeks. 2013. The promise and peril of real-time corrections to political misperceptions. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM pp. 1047–1058.

Garrett, R Kelly, Erik C Nisbet and Emily K Lynch. 2013. "Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory." *Journal of Communication* 63(4):617–637.

Geer, John G. 2008. *In defense of negativity: Attack ads in presidential campaigns*. University of Chicago Press.

Gentzkow, Matthew. 2006. "Television and voter turnout." *The Quarterly Journal of Economics* pp. 931–972.

Gentzkow, Matthew and Jesse M Shapiro. 2008. "Competition and Truth in the Market for News." *The Journal of Economic Perspectives* 22(2):133–154.

Gentzkow, Matthew and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78(1):35–71.

Gentzkow, Matthew, Jesse M Shapiro and Matt Taddy. 2015. "Measuring polarization in high-dimensional data: Method and application to congressional speech.".

Gentzkow, Matthew, Jesse M Shapiro and Michael Sinkinson. 2014. "Competition and ideological diversity: Historical evidence from us newspapers." *The American Economic Review* 104(10):3073–3114.

Geoffray, Marie Laure. 2014. "Channelling Protest in Illiberal Regimes: The Cuban Case since the Fall of the Berlin Wall." *Journal of Civil Society* 10(3):223–238.

Gerbaudo, Paolo. 2012. *Tweets and the streets: Social media and contemporary activism*. Pluto Press.

Gerber, Alan and Donald P Green. 1998. "Rational learning and partisan attitudes." *American journal of political science* pp. 794–818.

Gervais, Bryan T. 2015. "Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment." *Journal of Information Technology & Politics* 12(2):167–185.

Goel, Sharad, Winter Mason and Duncan J Watts. 2010. "Real and perceived attitude agreement in social networks." *Journal of personality and social psychology* 99(4):611.

Golbeck, Jennifer, Justin M Grimes and Anthony Rogers. 2010. "Twitter use by the US Congress." *Journal of the American Society for Information Science and Technology* 61(8):1612–1621.

Gott, Richard. 2011. *Hugo Chávez and the Bolivarian revolution*. Verso Books.

Green, Donald, Bradley Palmquist and Eric Schickler. 2002. "Partisan hearts and minds.".

Greenwood, S, A Perrin and M Duggan. 2016. "Social Media Update 2016." *Washington, DC: Pew Internet & American Life Project. Retrieved November* 27:2016.

Greitens, Sheena Chestnut. 2013. "Authoritarianism Online: What Can We Learn from Internet Data in Nondemocracies?" *PS: Political Science & Politics* 46(02):262–270.

Griffiths, Thomas L and Mark Steyvers. 2004. "Finding scientific topics." *Proceedings of the National academy of Sciences of the United States of America* 101(Suppl 1):5228–5235.

Gross, Terry. 2016. "Harassed On Twitter: 'People Need To Know The Reality Of What It's Like Out There'." *NPR* October 26, 2016.

Grossmann, Matt and David A Hopkins. 2016. *Asymmetric Politics: Ideological Republicans and Group Interest Democrats*. Oxford University Press.

Guess, Andrew. 2016. "Media Choice and Moderation:Evidence from Online Tracking Data." *Working Paper* .

Guess, Andrew, Brendan Nyhan and Jason Reifler. 2017. "Inside the Fake News Bubble? Consumption of online fake news in the 2016 U.S. election." *Working Paper* .

Guess, Andrew M. 2015. "Measure for Measure: An Experimental Test of Online Political

Media Exposure." *Political Analysis* 23(1):59–75.

Gulati, Jeff and Christine B Williams. 2010. "Communicating with constituents in 140 characters or less: Twitter and the diffusion of technology innovation in the United States Congress." *Available at SSRN 1628247* .

Gulker, Jill E, Aimee Y Mark and Margo J Monteith. 2013. "Confronting prejudice: The who, what, and why of confrontation effectiveness." *Social Influence* 8(4):280–293.

Gunitsky, Seva. 2015. "Corrupting the Cyber-Commons: Social Media as a Tool of Autocratic Stability." doi:10.1017/S1537592714003120.

Haidt, Jonathan. 2001. "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." *Psychological review* 108(4):814.

Haidt, Jonathan. 2012. "The righteous mind: Why good people are divided by politics and religion.".

Haidt, Jonathan. 2016. Why social media is terrible for multiethnic democracies. Technical report Vox.

Halpern, Daniel and Jennifer Gibbs. 2013. "Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression." *Computers in Human Behavior* 29(3):1159–1168.

Hamilton, James. 2004. *All the news that's fit to sell: How the market transforms information into news*. Princeton University Press.

Hanitzsch, Thomas, Maria Anikina, Rosa Berganza, Incilay Cangoz, Mihai Coman, Basyouni Hamada, Folker Hanusch, Christopher D Karadjov, Claudia Mellado, Sonia Virginia Moreira et al. 2010. "Modeling perceived influences on journalism: Evidence from a cross-national survey of journalists." *Journalism & Mass Communication Quarterly* 87(1):5–22.

Harfoush, Rahaf. 2009. *Yes We Did! An inside look at how social media built the Obama brand*. New Riders.

Hargittai, Eszter. 2001. "Second-level digital divide: mapping differences in people's online skills." *arXiv preprint cs/0109068* .

Hargittai, Eszter. 2005. "Survey measures of web-oriented digital literacy." *Social science computer review* 23(3):371–379.

Harrison, Brian F and Melissa R Michelson. 2012. "Not that there's anything wrong with

that: The effect of personalized appeals on marriage equality campaigns." *Political Behavior* 34(2):325–344.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman and R Tibshirani. 2009. *The elements of statistical learning*. Vol. 2 Springer.

Henson, Billy, Bradford W Reyns and Bonnie S Fisher. 2013. "Fear of crime online? Examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization." *Journal of Contemporary Criminal Justice* p. 1043986213507403.

Hilbe, Joseph M. 2008. "Brief overview on interpreting count model risk ratios: An addendum to negative binomial regression.".

Hindman, Matthew. 2008. *The myth of digital democracy*. Princeton University Press.

Hinduja, Sameer and Justin W Patchin. 2007. "Offline consequences of online victimization: School violence and delinquency." *Journal of school violence* 6(3):89–112.

Holbrook, Thomas M. 1999. "Political learning from presidential debates." *Political Behavior* 21(1):67–89.

Hong, Liangjie and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*. ACM pp. 80–88.

Hooghe, Marc and Ruth Dassonneville. 2011. "The effects of civic education on political knowledge. A two year panel survey among Belgian adolescents." *Educational Assessment, Evaluation and Accountability* 23(4):321–339.

Hope, Christopher. 2015. "And they're off: the 2015 general election campaign officially starts this Friday." *Telegraph UK* .

Hornik, Kurt and Bettina Grün. 2011. "topicmodels: An R package for fitting topic models." *Journal of Statistical Software* 40(13):1–30.

Hosseinmardi, Homa, Rahat Ibn Rafiq, Shaosong Li, Zhili Yang, Richard Han, Shivakant Mishra and Qin Lv. 2014. "A Comparison of Common Users across Instagram and Ask. fm to Better Understand Cyberbullying." *arXiv preprint arXiv:1408.4882* .

Howard, Philip N, Aiden Duffy, Deen Freelon, Muzammil Hussain, Will Mari and Marwa Mazaid. 2011. "Opening closed regimes: what was the role of social media during the Arab Spring?".

Howard, Philip N and Muzammil M Hussain. 2011. "The role of digital media." *Journal of*

Democracy 22(3):35–48.

Howard, Philip N and Muzammil M Hussain. 2013. *Democracy's Fourth Wave?: Digital Media and the Arab Spring*. Oxford University Press.

Howard, Philip N, Sheetal D Agarwal and Muzammil M Hussain. 2011. "When do states disconnect their digital networks? Regime responses to the political uses of social media." *The Communication Review* 14(3):216–232.

Hsu, Chien-leng and Han Woo Park. 2012. "Mapping online social networks of Korean politicians." *Government Information Quarterly* 29(2):169–181.

Huang, Haifeng. 2010. "Electoral Competition When Some Candidates Lie and Others Pander." *Journal of Theoretical Politics* 22(3):333–358.

Huber, Gregory A and Neil Malhotra. 2017. "Political homophily in social relationships: Evidence from online dating behavior." *The Journal of Politics* 79(1):269–283.

Huber, Gregory and Neil Malhotra. 2013. Dimensions of political homophily: Isolating choice homophily along political characteristics. In *American Political Science Association annual meeting, New Orleans, LA*.

Huddy, Leonie, Lilliana Mason and Lene Aarøe. 2015. "Expressive partisanship: Campaign involvement, political emotion, and partisan identity." *American Political Science Review* 109(01):1–17.

Huff, Connor and Dustin Tingley. 2015. ""Who Are These People?" Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents." *Research and Politics* 2(1):1–12.

Hwang, Hyunseo, Youngju Kim and Catherine U Huh. 2014. "Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation." *Journal of Broadcasting & Electronic Media* 58(4):621–633.

Hyde, Susan and Nikolay Marinov. 2009. "National Elections Across Democracy and Autocracy: Putting the 'Competitive'into Competitive Authoritarianism." *Unpublished Manuscript* .

Iyengar, Shanto. 1987. "Television news and citizens' explanations of national affairs." *American Political Science Review* 81(03):815–831.

Iyengar, Shanto and Donald R Kinder. 1987. "News that matters: Agenda-setting and priming in a television age." *News that Matters: Agenda-Setting and Priming in a Television*

Age .

Iyengar, Shanto, Gaurav Sood and Yphtach Lelkes. 2012. "Affect, not ideology a social identity perspective on polarization." *Public opinion quarterly* 76(3):405–431.

Iyengar, Shanto and Kyu S Hahn. 2009. "Red media, blue media: Evidence of ideological selectivity in media use." *Journal of Communication* 59(1):19–39.

Iyengar, Shanto, Kyu S Hahn, Heinz Bonfadelli and Mirko Marr. 2009. ""Dark Areas of Ignorance" Revisited Comparing International Affairs Knowledge in Switzerland and the United States." *Communication Research* 36(3):341–358.

Iyengar, Shanto and Sean J Westwood. 2015. "Fear and loathing across party lines: New evidence on group polarization." *American Journal of Political Science* 59(3):690–707.

Jerit, Jennifer and Jason Barabas. 2012. "Partisan perceptual bias and the information environment." *The Journal of Politics* 74(03):672–684.

Jerit, Jennifer, Jason Barabas and Toby Bolsen. 2006. "Citizens, knowledge, and the information environment." *American Journal of Political Science* 50(2):266–282.

Jerit, Jennifer, Jason Barabas, William Pollock, Susan Banducci, Daniel Stevens and Martijn Schoonvelde. 2016. "Manipulated vs. measured: Using an experimental benchmark to investigate the performance of self-reported media exposure." *Communication Methods and Measures* 10(2-3):99–114.

Jones, Jason J, Robert M Bond, Eytan Bakshy, Dean Eckles and James H Fowler. 2017. "Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 US presidential election." *PloS one* 12(4):e0173851.

Jun, Youjung, Rachel Meng and Gita Venkataramani Johar. 2017. "Perceived social presence reduces fact-checking." *Proceedings of the National Academy of Sciences* p. 201700175.

Jungherr, Andreas. 2010. "Twitter in politics: lessons learned during the german superwahljahr 2009." *CHI 2010* pp. 10–15.

Kam, Cindy D and Donald R Kinder. 2012. "Ethnocentrism as a short-term force in the 2008 American presidential election." *American Journal of Political Science* 56(2):326–340.

Kankaraš, Miloš, Guillermo Montt, Marco Paccagnella, Glenda Quintini and William Thorn. 2016. "Skills Matter: Further Results from the Survey of Adult Skills. OECD Skills Studies." *OECD Publishing* .

Karlsen, Rune and Eli Skogerbø. 2013. "Candidate campaigning in parliamentary systems Individualized vs. localized campaigning." *Party Politics* p. 1354068813487103.

Karpf, David. 2012*a*. *The MoveOn effect: The unexpected transformation of American political advocacy*. Oxford University Press.

Karpf, David. 2012*b*. "Social science research methods in Internet time." *Information, Communication & Society* 15(5):639–661.

Kennedy, M Alexis and Melanie A Taylor. 2010. "Online harassment and victimization of college students." *Justice Policy Journal* 7(1).

Khondker, Habibul Haque. 2011. "Role of the new media in the Arab Spring." *Globalizations* 8(5):675–679.

Kiesler, Sara, Jane Siegel and Timothy W McGuire. 1984. "Social psychological aspects of computer-mediated communication." *American psychologist* 39(10):1123.

Kiesler, Sara and Lee Sproull. 1992. "Group decision making and communication technology." *Organizational behavior and human decision processes* 52(1):96–123.

King, Gary, Jennifer Pan and Margaret E Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American Political Science Review* 107(02):326–343.

Klar, Samara and Yotam Shmargad. 2016. "Diverse Social Networks and Political Information Proliferation." *Presentation at NYU CESS Conference* .

Krafft, Peter M, Michael Macy and Alex Pentland. 2016. "Bots as Virtual Confederates: Design and Ethics." *arXiv preprint arXiv:1611.00447* .

Kruger, Justin, Nicholas Epley, Jason Parker and Zhi-Wen Ng. 2005. "Egocentrism over e-mail: Can we communicate as well as we think?" *Journal of personality and social psychology* 89(6):925.

Kuklinski, James H, Paul J Quirk, Jennifer Jerit, David Schwieder and Robert F Rich. 2000. "Misinformation and the currency of democratic citizenship." *Journal of Politics* 62(3):790–816.

Kuran, Timur. 1991. "Now out of never: The element of surprise in the East European revolution of 1989." *World politics* 44(01):7–48.

LaCour, Michael J and Lynn Vavreck. 2014. "Improving media measurement: Evidence from

the field." *Political Communication* 31(3):408–420.

Ladd, Jonathan M. 2011. *Why Americans hate the media and how it matters*. Princeton University Press.

Lapowsky, Issie. N.d. "Ev Williams on Twitter's Early Years.".

Lariscy, Ruthann Weaver, Elizabeth Johnson Avery, Kaye D Sweetser and Pauline Howes. 2009. "An examination of the role of online social media in journalists' source mix." *Public relations review* 35(3):314–316.

Lau, Richard R and David P Redlawsk. 2001. "Advantages and disadvantages of cognitive heuristics in political decision making." *American Journal of Political Science* pp. 951–971.

Lauderdale, Ben. 2015. "What We Got Wrong In Our 2015 U.K. General Election Model." *fivethirtyeight.com* .

Lawrence, Regina G and Amber E Boydstun. 2017. "What We Should Really Be Asking About Media Attention to Trump." *Political Communication* 34(1):150–153.

Lea, Martin and Russell Spears. 1991. "Computer-mediated communication, de-individuation and group decision-making." *International Journal of Man-Machine Studies* 34(2):283–301.

Leeper, Thomas J. 2016. "How does treatment self-selection affect inferences about political communication?" *Journal of Experimental Political Science* .

Leighley, Jan E and Jonathan Nagler. 2013. *Who Votes Now?: Demographics, Issues, Inequality, and Turnout in the United States*. Princeton University Press.

Lelkes, Yphtach, Gaurav Sood and Shanto Iyengar. 2015. "The hostile audience: The effect of access to broadband Internet on partisan affect." *American Journal of Political Science* .

Lelkes, Yphtach, Gaurav Sood and Shanto Iyengar. 2017. "The hostile audience: The effect of access to broadband internet on partisan affect." *American Journal of Political Science* 61(1):5–20.

Leshchenko, Sergii. 2014. "The Maidan and Beyond: The Media's Role." *Journal of Democracy* 25(3).

Levendusky, Matthew and Neil Malhotra. 2016. "Does media coverage of partisan polarization affect political attitudes?" *Political Communication* 33(2):283–301.

Lewis-Beck, Michael S and Martin Paldam. 2000. "Economic voting: an introduction." *Electoral studies* 19(2):113–121.

Lewis, Paul and Erin McCormick. 2018. *How an ex-YouTube insider investigated its secret algorithm*. The Guardian.

Lijphart, Arend. 1997. "Unequal participation: Democracy's unresolved dilemma presidential address, American Political Science Association, 1996." *American political science review* 91(01):1–14.

Little, Andrew T. 2012. "Elections, fraud, and election monitoring in the shadow of revolution." *Quarterly Journal of Political Science* 7(3):249–283.

Little, Andrew T. 2014. "Communication Technology and Protest.".

Livne, Avishay, Matthew P Simmons, Eytan Adar and Lada A Adamic. 2011. "The Party Is Over Here: Structure and Content in the 2010 Election.".

Lodge, Milton and Charles S Taber. 2013. *The rationalizing voter*. Cambridge University Press.

Lorentzen, Peter L. 2013. "Regularizing rioting: Permitting public protest in an authoritarian regime." *Quarterly Journal of Political Science* 8(2):127–158.

Lorenzo-Rodriguez, Javier. 2016. Who Tweets, About What and To Whom? In *Unpublished Manuscript*.

Lotan, Gilad, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce et al. 2011. "The Arab Spring— the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions." *International Journal of Communication* 5:31.

Lukianoff, Greg and Jonathan Haidt. 2015. "The coddling of the American mind.".

Lupia, Arthur. 1994. "Shortcuts versus encyclopedias: information and voting behavior in California insurance reform elections." *American Political Science Review* 88(01):63–76.

Magaloni, Beatriz and Ruth Kricheli. 2010. "Political order and one-party rule." *Annual Review of Political Science* 13:123–143.

Mainwaring, Scott. 2012. "From representative democracy to participatory competitive authoritarianism: Hugo Chavez and Venezuelan politics." *Perspectives on Politics* 10(04):955–967.

Malesky, Edmund and Paul Schuler. 2010. "Nodding or needling: analyzing delegate responsiveness in an authoritarian parliament." *American Political Science Review* 104(03):482–502.

Mann, Christopher B. 2010. "Is there backlash to social pressure? A large-scale field experiment on voter mobilization." *Political Behavior* 32(3):387–407.

Mantilla, Karla. 2013. "Gendertrolling: Misogyny Adapts to New Media." *Feminist Studies* pp. 563–570.

Martin, Gregory J and Ali Yurukoglu. 2014. Bias in cable news: Real effects and polarization. Technical report National Bureau of Economic Research.

Martin, Stephan. 2000. "The theory of contestable markets." *Bulletin of Economic Research* 37(1):65–68.

Mason, Lilliana. 2015. ""I disrespectfully agree": the differential effects of partisan sorting on social and issue polarization." *American Journal of Political Science* 59(1):128–145.

Mason, Lilliana. 2016. "A Cross-Cutting Calm How Social Sorting Drives Affective Polarization." *Public Opinion Quarterly* p. nfw001.

McCombs, Maxwell E and Donald L Shaw. 1972. "The agenda-setting function of mass media." *Public opinion quarterly* 36(2):176–187.

Mehrotra, Rishabh, Scott Sanner, Wray Buntine and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM pp. 889–892.

Meirowitz, Adam and Joshua A Tucker. 2013. "People Power or a One-Shot Deal? A Dynamic Model of Protest." *American Journal of Political Science* 57(2):478–490.

Messing, Solomon and Sean J Westwood. 2012. "Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online." *Communication Research* p. 0093650212466406.

Metzger, M. M., D. Penfold-Brown, P. Barbera, R. Bonneau, J. Jost, J. Nagler and J. Tucker. 2014. "Dynamics of influence in online protest networks: Evidence from the 2013 Turkish protests." *Paper presented at the annual meeting of the Midwest Political Science Association* .

Milner, Ryan M. 2013. "FCJ-156 Hacking the Social: Internet Memes, Identity Antagonism,

and the Logic of Lulz." *The Fibreculture Journal* (22 2013: Trolls and The Negative Space of the Internet).

Moor, Peter J. 2007. "Conforming to the flaming norm in the online commenting situation.".

Moor, Peter J, Ard Heuvelman and Ria Verleur. 2010. "Flaming on youtube." *Computers in Human Behavior* 26(6):1536–1546.

Morozov, Evgeny. 2011. "Whither Internet Control?" *Journal of Democracy* 22(2):62–74.

Morris, Jonathan S. 2005. "The Fox news factor." *The Harvard International Journal of Press/Politics* 10(3):56–79.

Morris, Stephen and Hyun Song Shin. 1998. "Unique equilibrium in a model of self-fulfilling currency attacks." *American Economic Review* pp. 587–597.

Mossberger, Karen, Caroline J Tolbert and Ramona S McNeal. 2007. *Digital citizenship: The Internet, society, and participation*. MIt Press.

Mosseri, Adam. 2016. "Building a Better News Feed for You." *Facebookl* .

Mungeam, Frank and Heather Crandall. 2011. "Commenting on the news: How the degree of anonymity affects flaming online.".

Munger, Kevin. 2016*a*. Interviews with Journalists. In *Unpublished Manuscript*.

Munger, Kevin. 2016*b*. Tweetment Effects on the Tweeted. In *Unpublished Manuscript*.

Munger, Kevin. 2016*c*. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* pp. 1–21.

Munger, Kevin. 2017*a*. "Don't@ Me: Experimentally Reducing Partisan Incivility on Twitter.".

Munger, Kevin. 2017*b*. The Economics of Online Journalism. In *Unpublished Manuscript*.

Munger, Kevin. 2017*c*. "Tweetment effects on the tweeted: Experimentally reducing racist harassment." *Political Behavior* 39(3):629–649.

Munger, Kevin and James Bisbee. 2016. The Trumpment Effect on the Trumped. In *Unpublished Manuscript*.

Munger, Kevin, Jonathan Ronen, Jonathan Nagler, Pat Egan and Joshua Tucker. 2016. The

Impact of Social Media Use on Voter Knowledge and Behavior in the 2015 UK Election: Evidence from a Panel Survey. In *Unpublished Manuscript*.

Munger, Kevin, Patrick Egan, Jonathan Nagler, Jonathan Ronen and Joshua A Tucker. 2016. "Learning (and Unlearning) from the Media and Political Parties: Evidence from the 2015 UK Election.".

Mutz, Diana C. 2015. *In-your-face politics: The consequences of uncivil media*. Princeton University Press.

Niculae, Vlad, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil and Jure Leskovec. 2015. QUOTUS: the structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee pp. 798–808.

Nielsen, Jakob. 2016. "The Distribution of Users' Computer Skills: Worse Than You Think." *Nielsen Norman Group* .

Nielsen, Rasmus Kleis and Cristian Vaccari. 2013. "Do people "like" politicians on Facebook? Not really. Large-scale direct candidate-to-voter online communication as an outlier phenomenon." *International Journal of Communication* 7:24.

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

Nyhan, Brendan, Jason Reifler, Sean Richey and Gary L Freed. 2014. "Effective messages in vaccine promotion: a randomized trial." *Pediatrics* 133(4):e835–e842.

Oates, Sarah. 2013. *Revolution stalled: The political limits of the Internet in the post-Soviet sphere*. Oxford University Press.

Omernick, Eli and Sara Owsley Sood. 2013. The Impact of Anonymity in Online Communities. In *Social Computing (SocialCom), 2013 International Conference on*. IEEE pp. 526–535.

O'Neill, Tom and Dawn Zinga. 2008. *Children's rights: Multidisciplinary approaches to participation and protection*. University of Toronto Press.

O'Reilly, Bill. 2003. *The no spin zone: Confrontations with the powerful and famous in America*. Three Rivers Press.

Paluck, Elizabeth Levy and Donald P Green. 2009. "Prejudice reduction: What works? A review and assessment of research and practice." *Annual review of psychology* 60:339–367.

Paluck, Elizabeth Levy, Hana Shepherd and Peter M Aronow. 2016. "Changing climates of conflict: A social network experiment in 56 schools." *Proceedings of the National Academy of Sciences* 113(3):566–571.

Panagopoulos, Costas. 2010. "Affect, social pressure and prosocial motivation: Field experimental evidence of the mobilizing effects of pride, shame and publicizing voting behavior." *Political Behavior* 32(3):369–386.

Papacharissi, Zizi. 2002. "The virtual sphere The internet as a public sphere." *New media & society* 4(1):9–27.

Papacharissi, Zizi. 2004. "Democracy online: Civility, politeness, and the democratic potential of online political discussion groups." *New Media & Society* 6(2):259–283.

Pearce, Katy. 2014. "Two can play at that game: Social media opportunities in Azerbaijan for government and opposition." *Demokratizatsiya* 22(1):39.

Perez, Valentina. 2014. "The Grim Reality of Venezuelan Protests." *Harvard Political Review* .

Pettigrew, Thomas F and Linda R Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of personality and social psychology* 90(5):751.

Pew Research Center. 2016. *Social Networking Use*. Pew.

Pew Research Center. 2017. *Mobile Fact Sheet*. Pew.

Phan, Xuan-Hieu, Le-Minh Nguyen and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*. ACM pp. 91–100.

Phillips, Whitney. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press.

Phillips, Whitney and R Milner. 2017. "The ambivalent internet: Mischief, oddity, and antagonism online." *Hoboken, NJ: Wiley* .

Piston, Spencer. 2010. "How explicit racial prejudice hurt Obama in the 2008 election." *Political Behavior* 32(4):431–451.

Plant, E Ashby and Patricia G Devine. 1998. "Internal and external motivation to respond without prejudice." *Journal of Personality and Social Psychology* 75(3):811.

Popper, Nathaniel. 2017. "Opioid Dealers Embrace the Dark Web to Send Deadly Drugs by Mail." *New York Times* June 10, 2017.

Posner, Sarah. 2016. "How Donald Trump's New Campaign Chief Created an Online Haven for White Nationalists." *Mother Jones* August 22, 2016.

Postmes, Tom and Russell Spears. 1998. "Deindividuation and antinormative behavior: A meta-analysis." *Psychological Bulletin* 123(3):238.

Postmes, Tom, Russell Spears, Khaled Sakhel and Daphne De Groot. 2001. "Social influence in computer-mediated communication: The effects of anonymity on group behavior." *Personality and Social Psychology Bulletin* 27(10):1243–1254.

Powell, G Bingham. 1986. "American voter turnout in comparative perspective." *American Political Science Review* 80(01):17–43.

Prior, Markus. 2005. "News vs. entertainment: How increasing media choice widens gaps in political knowledge and turnout." *American Journal of Political Science* 49(3):577–592.

Prior, Markus. 2007. *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

Prior, Markus. 2009*a*. "The immensely inflated news audience: Assessing bias in self-reported news exposure." *Public Opinion Quarterly* p. nfp002.

Prior, Markus. 2009*b*. "Improving media effects research through better measurement of news exposure." *The Journal of Politics* 71(03):893–908.

Prior, Markus. 2012. "Who watches presidential debates? Measurement problems in campaign effects research." *Public Opinion Quarterly* 76(2):350–363.

Prior, Markus. 2013*a*. "The challenge of measuring media exposure: Reply to Dilliplane, Goldman, and Mutz." *Political Communication* 30(4):620–634.

Prior, Markus. 2013*b*. "Media and political polarization." *Annual Review of Political Science* 16:101–127.

Prior, Markus and Arthur Lupia. 2008. "Money, time, and political knowledge: Distinguishing quick recall and political learning skills." *American Journal of Political Science* 52(1):169–183.

Rahimi, Babak. 2011. "The agonistic social media: cyberspace in the formation of dissent and consolidation of state power in postelection Iran." *The Communication Review* 14(3):158–

178.

Ramey, Adam, Jonathan Klingler and Gary E Hollibaugh. 2014. "More than a Feeling: Personality and Congressional Behavior." *Available at SSRN* .

Rand, David G, Thomas Pfeiffer, Anna Dreber, Rachel W Sheketoff, Nils C Wernerfelt and Yochai Benkler. 2009. "Dynamic remodeling of in-group bias during the 2008 presidential election." *Proceedings of the National Academy of Sciences* 106(15):6187–6191.

Rao, Delip, David Yarowsky, Abhishek Shreevats and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM pp. 37–44.

Rasinski, Heather M and Alexander M Czopp. 2010. "The effect of target status on witnesses' reactions to confrontations of bias." *Basic and Applied Social Psychology* 32(1):8–16.

Redlawsk, David P. 2002. "Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making." *The Journal of Politics* 64(04):1021–1044.

Reicher, Stephen D, Russell Spears and Tom Postmes. 1995. "A social identity model of deindividuation phenomena." *European review of social psychology* 6(1):161–198.

Remnick, David. 2016. Obama reckons with a Trump presidency. Technical report The New Yorker.

Reuter, Ora John and David Szakonyi. 2015. "Online social media and political awareness in authoritarian regimes." *British Journal of Political Science* 45(01):29–51.

Riker, William H. 1986. *The art of political manipulation*. Vol. 587 Yale University Press.

Riker, William H, Randall L Calvert and Rick K Wilson. 1996. *The strategy of rhetoric: Campaigning for the American Constitution*. Yale University Press.

Roberts, Margaret E, Brandon M Stewart and Dustin Tingley. 2014. "stm: R package for structural topic models." *R package version 0.6* 1.

Roberts, Margaret E, Brandon M Stewart and Edoardo Airoldi. 2015. "A model of text for experimentation in the social sciences." *Unpublished manuscript* .

Robertson, Graeme. 2011. "The politics of protest in hybrid regimes." *Managing dissent in post-communist Russia* .

Roeckelein, Jon E. 1998. *Dictionary of theories, laws, and concepts in psychology*. Greenwood Publishing Group.

Ronson, Jon. 2016. *So you've been publicly shamed*. Riverhead Books (Hardcover).

Rozenas, Arturas. 2010. Forced Consent: Information and Power in Non-Democratic Elections. In *APSA 2010 Annual Meeting Paper*.

Sanders, Lynn M. 1997. "Against deliberation." *Political theory* 25(3):347–376.

Sanders, Sam. 2017. "Upworthy Was One Of The Hottest Sites Ever. You Won't Believe What Happened Next." *NPR* June 20, 2017.

Sanovich, Sergey, Denis Stukal, Duncan Penfold-Brown and Joshua Tucker. 2015. "Turning the Virtual Tables: Government Strategies for Addressing Online Opposition with an Application to Russia.".

Scharkow, Michael. 2016. "The Accuracy of Self-Reported Internet Use—A Validation Study Using Client Log Data." *Communication Methods and Measures* 10(1):13–27.

Schattschneider, EE. 1960. "The Semisovereign People (New York: Holt, Rinehart and Winston, 1960)." *SchattschneiderThe Semi-Sovereign People1960* .

Schmierbach, Mike and Anne Oeldorf-Hirsch. 2012. "A little bird told me, so I didn't believe it: Twitter, credibility, and issue perceptions." *Communication Quarterly* 60(3):317–337.

Searles, Kathleen and Johanna Dunaway. 2017. "News and Information Loss in the Mobile Era." *Working Paper* .

Searles, Kathleen, Mingxiao Sui, Paul Newly and Johanna Dunaway. 2017. The Limits of Digital Citizenship: Constraints on News Consumption and Recall in the Mobile Setting. In *Unpublished Manuscript*.

Settle, Jaime. Forthcoming. *Newspaper to News Feed: How the Social Communication of Politics Affectively Polarizes the American Public*.

Shannon, Claude. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* pp. 379–423.

Shapiro, Robert Y. 2011. "Public opinion and American democracy." *Public Opinion Quarterly* 75(5):982–1017.

Shepherd, Hana and Elizabeth Levy Paluck. 2015*a*. "Stopping the Drama." *Social Psychology*

Quarterly 78(2):173–193.

Shepherd, Hana and Elizabeth Levy Paluck. 2015*b*. "Stopping the Drama Gendered Influence in a Network Field Experiment." *Social Psychology Quarterly* 78(2):173–193.

Sherif, Muzafer and Carolyn W Sherif. 1953. "Groups in harmony and tension; an integration of studies of intergroup relations.".

Shirky, Clay. 2008. *Here comes everybody: The power of organizing without organizations*. Penguin.

Sievert, Carson and Kenneth E Shirley. 2014. "LDAvis: A method for visualizing and interpreting topics.".

Silverman, Craig. 2016. This Analysis Shows How Fake Election News Stories Outperformed Real News On Facebook. Technical report Buzzfeed.

Smith, Aaron and Monica Anderson. 2018. *Social Media Use in 2018*. Pew.

Sood, Sara, Judd Antin and Elizabeth Churchill. 2012. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM pp. 1481–1490.

Stangor, Charles, Gretchen B Sechrist and John T Jost. 2001. "Changing racial beliefs by providing consensus information." *Personality and Social Psychology Bulletin* 27(4):486–496.

Stephens-Davidowitz, Seth. 2014. "The cost of racial animus on a black candidate: Evidence using Google search data." *Journal of Public Economics* 118:26–40.

Stephens-Davidowitz, Seth I. 2012. "The effects of racial animus on a black presidential candidate: using Google search data to find what surveys miss." *Available at SSRN 2050673* .

Stimson, James A. 2015. *Tides of consent: How public opinion shapes American politics*. Cambridge University Press.

Stringhini, Gianluca, Manuel Egele, Christopher Kruegel and Giovanni Vigna. 2012. Poultry markets: on the underground economy of twitter followers. In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*. ACM pp. 1–6.

Stroud, Natalie Jomini. 2008. "Media use and political predispositions: Revisiting the concept of selective exposure." *Political Behavior* 30(3):341–366.

Stroud, Natalie Jomini. 2011. *Niche news: The politics of news choice*. Oxford University Press on Demand.

Stroud, Natalie Jomini, Joshua M Scacco, Ashley Muddiman and Alexander L Curry. 2014. "Changing deliberative norms on news organizations' Facebook sites." *Journal of Computer-Mediated Communication* 20(2):188–203.

Sydell, Laura. 2016. "We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned." *NPR.com* .

Taber, Charles S and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American Journal of Political Science* 50(3):755–769.

Tajfel, Henri and John C Turner. 1979. "An integrative theory of intergroup conflict." *The social psychology of intergroup relations* 33(47):74.

Theocharis, Yannis, Pablo Barberá, Zoltan Fazekas and Sebastian Adrian Popa. 2015. "A Bad Workman Blames His Tweets? The Consequences of Citizens' Uncivil Twitter Use When Interacting with Party Candidates." *The Consequences of Citizens' Uncivil Twitter Use When Interacting with Party Candidates (September 5, 2015)* .

Tillman, Erik R. 2012. "Support for the euro, political knowledge, and voting behavior in the 2001 and 2005 UK general elections." *European Union Politics* 13(3):367–389.

Trippi, Joe. 2004. "The revolution will not be televised." *CAMPAIGNS AND ELECTIONS* 25(8):44–44.

Truman, David Bicknell et al. 1971. *The governmental process*. Alfred A. Knopf New York.

Tucker, Josh, Megan Metzger, Duncan Penfold-Brown, Richard Bonneau, John Jost and Johnathan Nagler. 2014. "Protest in the Age of Social Media: Ukraine's Euromaidan." *Carnegie Corporation* .

Tucker, Joshua A. 2007. "Enough! Electoral fraud, collective action problems, and post-communist colored revolutions." *Perspectives on Politics* 5(03):535–551.

Tucker, Joshua A and Radoslaw Markowski. 2007. Subjective vs. Objective Proximity in Poland: New Directions for the Empirical Study of Political Representation. 2007 Annual Meeting of the American Political Science Association.

Tufekci, Zeynep. 2014. "Social movements and governments in the digital age: Evaluating a complex landscape." *Journal of International Affairs* 68(1):1.

Tufekci, Zeynep. 2017. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Tufekci, Zeynep and Christopher Wilson. 2012. "Social media and the decision to participate in political protest: Observations from Tahrir Square." *Journal of Communication* 62(2):363–379.

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

Vaccari, Cristian and Rasmus Kleis Nielsen. 2013. "What Drives Politicians' Online Popularity? An Analysis of the 2010 US Midterm Elections." *Journal of Information Technology & Politics* 10(2):208–222.

Valenzuela, Sebastián, Yonghwan Kim and Homero Gil de Zúñiga. 2012. "Social networks that matter: Exploring the role of political discussion for online political participation." *International journal of public opinion research* 24(2):163–184.

Van Deursen, Alexander JAM, Ellen J Helsper and Rebecca Eynon. 2016. "Development and validation of the Internet Skills Scale (ISS)." *Information, Communication & Society* 19(6):804–823.

Vandebosch, Heidi and Katrien Van Cleemput. 2009. "Cyberbullying among youngsters: Profiles of bullies and victims." *New media & society* 11(8):1349–1371.

Walther, Joseph B. 1996. "Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction." *Communication research* 23(1):3–43.

Weaver, Matthew. 2015. "Nick Clegg's tuition fees 'debacle' undermined trust, says Norman Lamb." *The Guardian* May 12.

Weaver, Vesla M. 2012. "The electoral consequences of skin color: The "hidden" side of race in politics." *Political Behavior* 34(1):159–192.

Whiteley, Paul, Harold D Clarke, David Sanders and Marianne C Stewart. 2013. *Affluence, Austerity and Electoral Change in Britain*. Cambridge University Press.

Willnat, Lars and David Hugh Weaver. 2014. *American Journalist in the Digital Age: Key Findings*.

Wilson, Robert E, Samuel D Gosling and Lindsay T Graham. 2012. "A review of Facebook research in the social sciences." *Perspectives on psychological science* 7(3):203–220.

Wojcieszak, Magdalena E and Diana C Mutz. 2009. "Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement?" *Journal of communication* 59(1):40–56.

Wulczyn, Ellery, Nithum Thain and Lucas Dixon. 2016. "Ex Machina: Personal Attacks Seen at Scale." *arXiv preprint arXiv:1610.08914* .

Wulczyn, Ellery, Nithum Thain and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee pp. 1391–1399.

Xu, Zhi and Sencun Zhu. 2010. Filtering offensive language in online communities using grammatical relations. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*.

Ybarra, Michele L, Danah Boyd, Josephine D Korchmaros and Jay Koby Oppenheim. 2012. "Defining and measuring cyberbullying within the larger context of bullying victimization." *Journal of Adolescent Health* 51(1):53–58.

Yeo, Sara K, Michael A Xenos, Dominique Brossard and Dietram A Scheufele. 2015. "Selecting Our Own Science How Communication Contexts and Individual Traits Shape Information Seeking." *The ANNALS of the American Academy of Political and Social Science* 658(1):172–191.

Yin, Dawei, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis and Lynne Edwards. 2009. "Detection of harassment on web 2.0." *Proceedings of the Content Analysis in the WEB* 2.

Zaller, John. 1992. *The nature and origins of mass opinion*. Cambridge university press.

Zitek, Emily M and Michelle R Hebl. 2007. "The role of social norm clarity in the influenced expression of prejudice over time." *Journal of Experimental Social Psychology* 43(6):867–876.