Original Research

The Dynamics of Political Incivility on Twitter

SAGE Open
April-June 2020: 1–15

DOI: 10.1177/2158244020919447
journals.sagepub.com/home/sgo
SAGE

Yannis Theocharis[1], Pablo Barberá[2], Zoltán Fazekas[3], and Sebastian Adrian Popa[4]

Abstract

Online incivility and harassment in political communication have become an important topic of concern among politicians, journalists, and academics. This study provides a descriptive account of uncivil interactions between citizens and politicians on Twitter. We develop a conceptual framework for understanding the dynamics of incivility at three distinct levels: macro (temporal), meso (contextual), and micro (individual). Using longitudinal data from the Twitter communication mentioning Members of Congress in the United States across a time span of over a year and relying on supervised machine learning methods and topic models, we offer new insights about the prevalence and dynamics of incivility toward legislators. We find that uncivil tweets represent consistently around 18% of all tweets mentioning legislators, but with spikes that correspond to controversial policy debates and political events. Although we find evidence of coordinated attacks, our analysis reveals that the use of uncivil language is common to a large number of users.

Keywords

politics, social media, incivility, machine learning, gender, topic models

> . . . And then I had no idea what was about to happen next. My Twitter feed basically exploded. [. . .] I began to see images, for example, of my youngest daughter, who we adopted from Ethiopia many years ago, who at the time was 7 years old—images of her in a gas chamber with a—Donald Trump in an SS uniform about to push the button to kill her.
>
> *—U.S. attorney David French (interview with NPR)*

> To see the attack of a pack on here check out my mentions 600 odd notifications talking about my rape in one night. I think twitter is dead
>
> *—Jess Phillips, Labour MP for Birmingham Yardley (on Twitter)*

Introduction

Politics has always been an arena of heated argumentation. Witty, caustic, ironic, and oftentimes vitriolic verbal exchanges are part of a discourse that defines political competition and delineates power relations between political actors. At the same time, such exchanges often provide signals about what is permissible in public discourse. A politician's choice of words, therefore, can as much restrain and reconciliate, as it can spread division and elevate the status of offensiveness from unacceptable to routine.

Operating in an arena where strongly caustic language and argumentation are common, one would expect that politicians would be unsurprised with (and resilient against) the outcome of the public's newly acquired capacity to address them directly via new media channels, such as through the microblogging platform Twitter. However as high as a politician's public approval may be, it is no secret that citizens are cynical toward politicians (Hay, 2007). After all, parties are the democratic institution that tends to be the least trusted of all others, in Europe as well as in the United States (Capella & Jamieson, 1997; Torcal & Montero, 2006). Thus, an anonymous platform for direct and publicly visible communication with one's representatives can be a natural channel for citizens to communicate some of this frustration in the form of heated remarks. But what if this communication entails more than just a bit of justified frustration expressed in cynical language? What if this interaction is uncivil and entails prejudiced, hurtful, discouraging, and damaging language that is unpreventable and constant?

[1]University of Bremen, Bremen, Germany
[2]University of Southern California, Los Angeles, USA
[3]Copenhagen Business School, Frederiksberg, Denmark
[4]Newcastle University, Newcastle upon Tyne, UK and MZES, University of Mannheim, Mannheim, Germany

Corresponding Author:
Pablo Barberá, University of Southern California, VKC 330, 3518 Trousdale Parkway, Los Angeles, CA 90089, USA.
Email: pbarbera@usc.edu

Online incivility and harassment in political communication have become an important topic of concern among politicians, journalists and academics. Although the phenomenon of rising incivility within the political arena (Mutz, 2015; Uslaner, 1993) and in relation to it (Funk, 2001; Sigelman & Bullock, 1991) has been widely discussed (Mutz & Reeves, 2005), social media communication has reinvigorated this debate by highlighting new aspects to consider due to platforms' varied affordances. Empowered by the interactive capacities of platforms like Twitter or Facebook, individual users (but also bots) can now directly and publicly address comments to their representatives under conditions of anonymity. The capacity of social media to strengthen the relationship between representatives and their constituents has long been seen as a potentially major advance toward a more inclusive public sphere and toward strengthening public deliberation (Coleman, 2005). Yet, the danger of incivility—and the further concern of its normalization—in such anonymous environments looms large (Coe et al., 2014; Sobieraj & Berry, 2011) and can have important consequences for democracy.

Previous research has found that incivility has strong negative effects on attitudes and behaviors at many different levels (Anderson et al., 2013; Gervais, 2015; Massaro & Stryker, 2012). First, exposure to incivility *between politicians* has been associated with the public's dissatisfaction with political institutions and negative attitudes toward politicians (Capella & Jamieson, 1997; Elving, 1994, but see Brooks & Geer, 2007). Second, exposure to online incivility *between citizens* in places like blogs and online forums can decrease open-mindedness, political trust, and efficacy (Borah, 2012) and polarize individuals' views on a topic (Anderson et al., 2013; Lyons & Veenstra, 2016). In addition, harassment directly aimed at individuals, especially minorities and vulnerable groups, tends to make them more anxious for their safety and demobilize them (Henson et al., 2013; Hinduja & Patchin, 2007; Munger, 2016). Third, in the least developed research area on uncivil interactions, that *between citizens and politicians*, evidence shows that although political candidates do occasionally engage in reciprocal engagement with citizens (Enli & Skogerbø, 2013; Tromble, 2018), those who make engaging use of Twitter are more likely to receive impolite and uncivil tweets than those who tend to simply "broadcast" (Theocharis et al., 2016).

In this study, we focus our attention on uncivil interactions between citizens and politicians on Twitter, one of the most popular social media platforms. Twitter constitutes an important new arena of engagement between representatives and their constituents, with major consequences for democracy as, for the first time, the two sides are able to engage in—and both benefit from—direct communication. At the same time, this is also an arena about which, aside from the media coverage of famous cases, we still know little regarding the phenomenon of incivility, its dynamics, and its mechanisms. There are, however, clear indications that it is becoming a worrying trend that needs attention (BBC, 2017; The Guardian, 2016; Halliday, 2012; Hayes, 2008), to the extent that Twitter has been forced to experiment with new measures to crack down on it (Pham, 2017).

Our goal is to offer a descriptive account of the extent to which uncivil behavior toward politicians is prevalent on Twitter and to offer an exploratory look at what explains variation over time and across politicians. We use longitudinal data from the Twitter communication mentioning members of the U.S. Congress across a time span of over a year. This allows us to observe uncivil behavior toward politicians as it unfolds, not only during heated events such as electoral campaigns but also during *quieter* periods in between elections, which may, however, be occasionally characterized by social or political events that lead to spikes in citizen and media attention and, potentially, uncivil communication.

We develop and test a set of expectations to explain incivility at three distinct levels—macro, meso, and micro: on the *macro* level, we are interested in the temporal dynamics of incivility; on the *meso* level, we study the role of contextual factors such as the type of issues that are part of a given day's agenda; finally, on the *micro* level, we are interested in the role played by individual prejudice, focusing in particular on the uncivil behavior and harassment aimed at women.

Our results show that the level of incivility addressed to Members of Congress on Twitter is relatively stable over time, around 18% of all tweets. This figure corroborates findings from other studies concerning the amount of incivility aimed at U.S. Senators (Rheault et al., 2019). However, we also observe spikes that correspond to moments of controversy about specific policy issues, such as health care, or political events, such as the White nationalist rally in Charlottesville, VA. Regarding who is responsible for the uncivil tweets, we find that up to 36% of all unique users in our data set sent at least one uncivil tweet, although again we find evidence of coordinated attacks at specific moments in time. Finally, regarding who is the target of incivility, we actually do not find large differences based on the gender of the legislator; in contrast, party and ideology are more important determinants of who is victim of harassment on Twitter.

Four Dynamic Elements of Incivility on Social Media

Extant literature shows that politicians, especially in the United States and Europe, have adopted social media widely, as tools that can strengthen communication with their constituents (Barberá & Zeitzoff, 2017; Bode & Dalrymple, 2016; Gulati & Williams, 2010, 2013; Heiss et al., 2018; Jungherr, 2016; Nulty et al., 2016; Popa et al., 2019; Williams & Gulati, 2010). Twitter, in particular, allows them to reach voters directly, bypassing middlemen and offering the opportunity to marketize themselves and move closer to voters by presenting a more personalized version of themselves (Enli & Skogerbø, 2013; Jungherr, 2016; Karlsen & Skogerbø, 2015). Although some politicians make interactive use of Twitter, just as with other social media platforms, in their

majority they tend to use the platform mostly during electoral campaigns and, in principle, as broadcasting rather than as an interactive tool (Graham et al., 2013, 2014; Jackson & Lilleker, 2011; Larsson, 2015; Williamson, 2010). This has been seen as an unfortunate state of affairs, as interactive use not only leads to more positive evaluations of politicians by citizens (Lee & Shin, 2012; Lyons & Veenstra, 2016) but can also be an opportunity for citizens to learn more about party platforms (Fazekas et al., 2018; Munger et al., 2016).

The one-directional use of social media by politicians has generally been theorized using supply-side explanations, including, for example, that politicians do not have sufficient time to engage, are not technologically up to speed, or are concerned about losing control of the message (Stromer-Galley, 2000; Ward & Lusoli, 2005). Yet the unique dynamics that emerge through citizen–elite interactions on social media—especially Twitter—open the door for other explanations focused on the newly acquired capacities of the demand side to directly address the supply side. In one of the few studies focusing on demand-side explanations, Tromble (2018) suggests that political elites may be failing to directly engage with citizens to avoid meeting hostility with hostility and because the risks of negative reciprocity are high. The important role of harassment and abuse by citizens as major preventive factors for politicians to engage with the public should not come as a surprise. Politicians have been the targets of abuse, threats, and violence long before the arrival of social media. As McLoughlin and Ward (2017) explain in their work on abuse of MPs in the British Parliament, politicians have traditionally attracted abuse due to their elevated public profile and the public's image of them as people with power, with explanations for these tendencies lying less with psychological factors (James et al., 2016) and more with factors such as disenchantment, boredom, or elements of a broader "trolling" subculture (Every-Palmer et al., 2015; Marwick & Lewis, 2017).

While the phenomenon is not new, social media and their (mis)affordances have amplified and, as public commentators have argued (Kamps, 2016), perhaps more dangerously from a democratic point of view, normalized it (Pelled et al., 2018). As misaffordances, here we consider (a) the combined *outcome* of making very slow steps in efficiently moderating the discussion on the side of social media companies, and (b) the *mode of communication encouraged by the platform*, which is low-cost and (potentially) anonymous, making emotional engagement easier than it would have been in face-to-face communication (Frijda, 1988; Suler, 2004) and, in many cases, penalty-free. Beyond innumerable anecdotal evidence about political harassment and trolling that led to political actors terminating their engagement on Twitter and a study on the broader level of abuse received by British MPs (McLoughlin & Ward, 2017), the theoretical assumption concerning the changing dynamics of politician interactions on Twitter due to uncivil attacks has remained largely untested in the literature. Although Munger (2016) and Theocharis et al. (2016) have established that the phenomenon *is* occurring frequently during periods of vibrant political activity, such as around elections—and that it affects politicians' communication patterns negatively—there is much less evidence on the magnitude of the phenomenon and of its different forms and consequences.

Defining (in)civility is a challenging endeavor due to the complexity of the concept (Herbst, 2010, p. 12). Most scholars understand incivility as a multidimensional concept. For example, Sapiro suggests three dimensions, one mapping on good character and virtue, another one on manners, and a final one to its communicative nature (Sapiro, 1999); Sydnor emphasizes tone and perceives it as a continuum "that ranges from the polite to insults to racial slurs and obscenities" (Sydnor, 2019, p. 9); Muddiman (2017) distinguishes between personal- and public-level incivility, whereby the first focuses on rudeness and emotions and the second on aspects such as failing to recognize the legitimacy of opposing views. Distinctions in terms of tone and morality have also been drawn by Papacharissi (2004), who differentiated between impoliteness and incivility early, with the second linking closely to intolerant behavior and thus with the idea of disrespect toward collective democratic traditions. In more recent contributions, it is this last aspect that makes incivility a threatening element to democratic pluralism and distinguishes it from incivility as a rhetorical act—a device that does not necessarily have harmful effects (Rossini, 2019).

As defining incivility is beyond the scope of this study, we follow Herbst (2010, p. 12) and adopt a definition that makes sense of the level and nature of our empirical work. Accordingly, we perceive incivility more broadly as a "disrespectful discourse that silences or derogates alternative views" (Jamieson et al., 2017, p. 206). Building on previous studies acknowledging the importance of context, as well as that of public versus private levels of incivility (Muddiman, 2017), we argue that to understand the occurrence of incivility in Twitter conversations between citizens and politicians—as well as its consequences for the second—it is essential to develop designs that allow us to observe such behavior from a bird's-eye view. Such a view can reveal the temporal and contextual levels at which such behavior occurs, as well as better capture uncivil conversational dynamics (such as virality) that can become mostly evident on social media platforms such as Twitter due to their distinct architecture. For example, incivility on Twitter may not be limited to an infrequent type of behavior exhibited sporadically by a few frustrated citizens wishing to speak their mind in a one-off, emotional impolite remark. Indeed, as the many cases of harassment on Twitter have shown—such as the hate, racist, and sexist campaigns toward Black actress Leslie Jones (Oluo, 2016), United Kingdom's Labour MP Tulip Siddiq (Saner, 2016), or Canada's Progressive Conservative politician Sandra Jensen (BBC, 2016), it may well be a coordinated and continuous effort from a group of people to

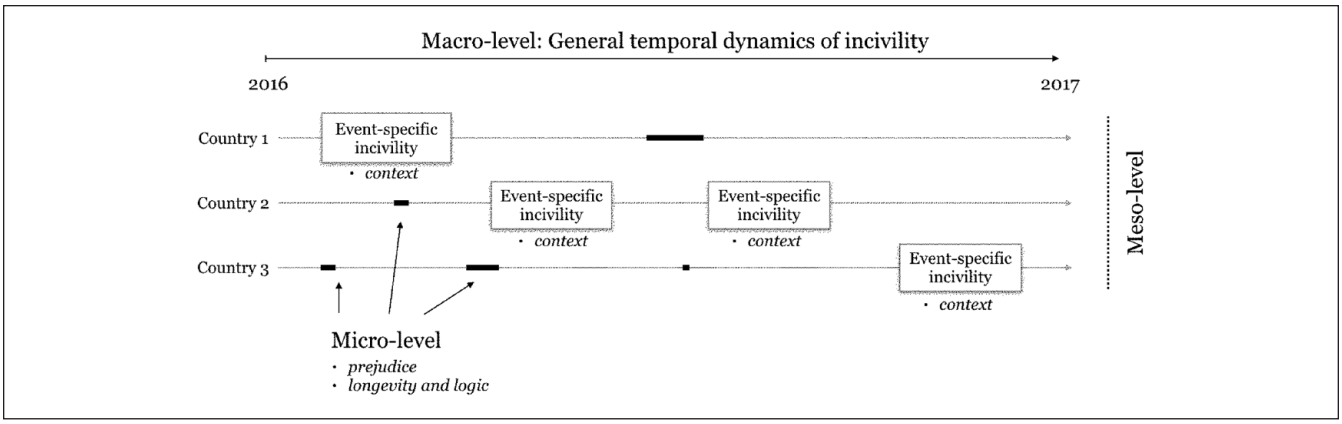

Figure 1. Conceptual framework: The three levels of incivility.

cause distress, smear, and undermine someone's public profile by continuously harassing her or by causing a cascade effect in which an impolite or uncivil remark may encourage others to join in.

Similarly, the many media reports about specific groups becoming more often than others targets of incivility (Saner, 2016) may be an indication that prejudice toward a group with specific characteristics may be a default trigger for incivility. These aspects of incivility can be found within the same communicative framework and can be systematically thought of as operating, and being spatially interlinked, on three levels (see Figure 1). We distinguish between the macro level, which allows us to observe how incivility unfolds across time; the meso level, which allows us to study the role of specific events or issues within the political agenda; and the more granular, micro level, which enables us to better understand who gets targeted and by whom. The different elements of incivility we analyze are represented schematically in Figure 1 and discussed separately in the following sections. While our study cannot speak directly to the effects of incivility as a result of these dynamics (e.g., demobilization), our conceptual framework makes the first step toward exploring the occurrence of incivility under different temporal and contextual conditions—as well as logics. As such, this framework can offer several points of departure for future research dealing both with the causes and the effects of the dynamics uncovered here.

Macro Level: Temporality

Temporality can be thought of as the highest level of our analysis. Scholarly work on Twitter use by European Union (EU) parliamentarians has shown that their Twitter communication is not characterized by permanence (Larsson, 2015). At the same time, studies focusing on online incivility (Coe et al., 2014) have shown that incivility tends to occur frequently and in response to interactive use on the part of politicians (Theocharis et al., 2016). Yet, previous work on political incivility on Twitter has been limited by the relatively short time periods examined, usually spanning no more than a month and on specific events. Studying abusive tweets aimed at U.K. parliamentarians for 2½ months, for example, McLoughlin and Ward (2017, p. 11) found evidence of significant groupings of abuse on particular days, concluding that abuse may not be a day-to-day occurrence but rather something fueled by outside factors. Indeed, while some of the existing approaches provide valuable information about the extent of uncivil interactions in specific cases, such as electoral campaigns, they focus on time periods in which political interest and discussion are heightened. As during such periods politicians tend to use Twitter more, patterns of incivility might differ from "quieter" periods in between elections. Our data set involves interactions between citizens and politicians for approximately a year, enabling us to study these nuances across lengthier time spans than any previous study, which allows us to examine also politically quiet and tense periods. We expect that, in general, incivility is not a constant state of affairs in the Twitter communication of politicians, but a phenomenon that peaks around highly mediated events which may not be limited to elections but may include scandals, crises, or terrorist attacks.

> **Hypothesis 1 (temporality):** The volume of uncivil tweets targeted at politicians will be significantly higher around highly communicative events (such as election campaigns, political scandals, and major legislative debates) than during uneventful periods.

Meso Level: Context

The hypothesis stated above opens up one further important question. Regardless of whether incivility peaks at certain times or whether it is a constant state of affairs, are there specific debates or events that lead to larger outbursts of vitriolic reactions? The interactive communication structure of Twitter, in combination with the status and individual

characteristics of the political actor tweeting, may imply a similar dynamic when it comes to how—and which—issues drive incivility. This means that incivility toward politicians may be conditioned by contextual factors not limited to electoral campaigns, but including topics such as corruption scandals, economic crises, terrorist attacks, speeches, vote on a specific legislation, and televised addresses or debates.

Previous research on uncivil comments in online newspaper articles by Coe et al. (2014) found that incivility is associated with contextual factors, such as the topic of the article. Specifically, "hard news" issues such as the economy, law, politics, taxes, and foreign affairs tend to receive far more uncivil comments than issues such as health, lifestyle, and journalism (Coe et al., 2014, p. 669). Similarly, research on Twitter communication shows that conversations about "hard" political issues are more polarized discussions (Barberá et al., 2015) which may also lead to, overall, more heated discussions.

The role of the context within which incivility takes place is located at the meso level of our conceptual framework, as it represents distinct, short- or long-lived events within a given broader temporal frame. Thus, building previous work by Coe et al. (2014) and Barberá et al. (2015), our second hypothesis is as follows:

> **Hypothesis 2 (context):** The volume of uncivil tweets targeted at politicians will be significantly higher in response to messages about "hard" topics such as economic and social issues.

Examining contextual factors driving incivility is important from two points of view. First, it can allow insights into which topics may be alerting politicians about possible forthcoming waves of incivility and how they modify their behavior accordingly (i.e., not engaging, or retreating after being harassed) to avoid a communication crisis. Second, it enables us to study whether high levels of incivility when engaging with users about a certain topic lead to discouragement from engaging with that topic in the future. Aside from the dangers embedded in the coarsening of democratic discourse and the normalization of incivility, such a development would imply the impoverishment of the debate through the avoidance of "unsafe" topics—a well-documented pitfall of choosing not to engage in conflict in democratic deliberation literature (Papacharissi, 2004). A direct result is that less information will be communicated to citizens via this avenue, and there will be more hesitance from the politicians to address potentially important issues.

Micro Level: Coordination and Prejudice

Although some have argued that minorities and other vulnerable populations are often subjected to consistent online harassment, aside from cases that have attracted media attention, there is little research on whether incivility toward a politician has a limited time span or whether it tends to persist across a longer period of time. Furthermore, if incivility toward politicians on Twitter is event-driven, then one might be led to believe that users who engage in this behavior are frustrated citizens who, motivated by the particular event, decide to channel their frustration on Twitter in the form of a one-off aspersion. If this is the case, then incivility will be a sporadic type of behavior, carried out occasionally by different users.

This would be consistent with some of the previous work on online forums by Coe et al. (2014), which examined whether incivility is the purview of frequent or occasional commentators and found the second to be more uncivil. Similarly, McLoughlin and Ward (2017) found that the users who had sent abusive tweets to British MPs in their data set were not "serial transgressors" (or groups of such) but rather that abuse was distributed across a large number of users (p. 16). Twitter's communication structure and affordances make it an ideal place for engaging in sporadically uncivil behavior. The possibility of remaining anonymous, along with the low-cost of targeting someone, makes it an appealing platform for a spontaneous, hit-and-run type of behavior encouraged possibly by the feeling of it being easier to hide in the crowd (Oz et al., 2018). Yet research on trolling subcultures has shown that organized trolling is a very real—and not new—phenomenon (Phillips, 2016), while much less attention has been paid to the more recent development of nonhuman (i.e., bot) abuse. Moving to the micro level with the aim to better understand the logic and longevity of incivility on the basis of who and how is engaging in this type of behavior, we hypothesize the following:

> **Hypothesis 3 (longevity and logic):** Uncivil tweeting is carried out by different users in a sporadic fashion.

Although the previous hypothesis reveals some of the mechanics of user-initiated incivility on Twitter, they convey little information about potential motives behind uncivil behavior toward politicians using the platform. Although it is unlikely that any approach can offer firm evidence as to what one's motives are for targeting a politician, studying the characteristics of the victims can provide information with regard to whether prejudice toward a subgroup with common characteristics might be one of the driving forces of incivility. Here, we focus on gender-based incivility, although this is only one of the many subgroups that incivility has been found to be directed toward, with ethnic, ideology-based, and racial harassment also being common (Munger, 2016).

Research on gendered coverage has found that electoral candidates are often portrayed in terms of long-standing gender stereotypes, and that gender differences have important consequences for voters' perceptions, especially of female contenders (Kittilson & Fridkin, 2008). In light of the bulk of previous work showing that media coverage further raises the barriers faced by female contenders seeking office (for an

overview, see Lühiste & Banducci, 2016), one would think that social media could be an empowering communication tool giving women a more independent and bias-free avenue of communication. Yet, a 2016 study commissioned by *The Guardian* showed that in the context of online news articles, women authors tend to be treated differently by commenters and receive more abuse, a tendency consistent with characteristics of a phenomenon Mantilla has referred to as "gendertrolling" (Mantilla, 2015). Similar findings were reported by Amnesty International (2018) which declared Twitter a toxic place for women. The extent of uncivil behavior directed toward specifically women parliamentarians—in particular in the form of harassment and online threats—was also confirmed by a study commissioned by the Inter-Parliamentary Union (2016), which corroborates previous research (Henson et al., 2013). Despite this evidence, we know little about the extent to which candidates' personal characteristics provide a default trigger of incivility in citizen–politician interactions.

Investigating whether prejudice is the basis of a uncivil behavior is particularly critical from a democratic point of view, as harassment on the basis of individual characteristics (e.g., gender, race) not only implies a stark disrespect toward collective democratic traditions (Papacharissi, 2004, p. 260), but it is also a type of behavior that is very difficult to deal with. At the same time, as other studies have shown, it is also a type of behavior that tends to have poisonous effects on the victims, including making them more anxious for their safety and demobilizing them (Hinduja & Patchin, 2007; Munger, 2016). Jess Philips, the Labour MP quoted at the beginning of this article, is just one of the many examples who, after being harassed on the basis of her gender, felt that she would be better off if she quit Twitter or if she reduced her general use of the platform. Based on these considerations, we hypothesize the following:

> **Hypothesis 4 (prejudice):** Female politicians are significantly more likely than men to receive uncivil comments.

Data Collection and Case Selection

We study incivility from the multilevel perspective derived from our conceptual framework by analyzing a large-scale longitudinal Twitter data collection that contains public responses to messages shared by Members of Congress in the United States.

To initiate our data collection, we relied on the list of Twitter accounts of Members of Congress elected to the 115th Congress (2017–2018) available in the *unitedstates* GitHub account. We then downloaded all tweets that mention any of these politicians' Twitter accounts directly from Twitter's Streaming API (application programming interface).[1]

Our full data set contains a total of 16,002,098 tweets mentioning a Member of Congress from October 17, 2016, until December 13, 2017. These tweets can be divided into two categories: 11,222,146 *replies* (tweets that directly respond to a tweet by a legislator and are labeled as such if they were seen on the website) and 4,779,952 *mentions* (tweets that explicitly mention a legislator's Twitter handle but are not a direct response to one of their tweets). We pool both types of tweets in our analysis, with the exception of our study of incivility in response to specific topics addressed by legislators, where we only use replies because those are the only type of tweets we can directly match to a specific statement by a legislator on Twitter. In addition, we also downloaded the 155,540 tweets by Members of Congress that received at least one reply during the same period, which we will use to determine which policy issues receive a more uncivil response.

This data set covers a long time span, which is ideal for our temporal investigation, while at the same time including a series of highly politicized events for our event-driven hypotheses, such as the presidential election in November 2016 or the beginning of the Russia investigation in May 2017. As a result, we believe it presents an accurate reflection of periods of both stability and heated political competition and allows us to conduct various investigations on the temporal dynamics of incivility from the macro to micro level.

Method

We test our four hypotheses using a variety of metrics computed using automated text analysis methods. As we explain in greater detail below, we detect incivility in tweets using supervised learning methods, which extrapolate human coding on a random sample of tweets to the entire data set by identifying words that tend to be associated with uncivil responses. To classify tweets by politicians into different types of topics, we instead use a topic model, which gives us greater flexibility in contexts where multiple topics may be relevant. We measure coordinated incivility campaigns by computing the degree to which uncivil tweets are concentrated on a small subset of users using the Gini index. Finally, our examination of who is the target of incivility relies on covariates at the legislator level.

Automatic Detection of Incivility in Social Media Posts

In our analysis, we apply supervised machine learning to detect incivility in users' tweets. This method relies on a training sample of 4,000 tweets, randomly sampled from our full data set, which were annotated manually by humans along our dimension of interest. For this labeling task, the coders were selected through CrowdFlower (now called

Figure Eight), an online platform for crowd-coding that has been found to yield accurate results with high intercoder reliability (Benoit et al., 2016).

For each of the tweets, we asked the coders to label it using a coding scheme inspired by the codebook developed by Theocharis et al. (2016). We consider the following two categories:

1. *Civil*: a tweet that adheres to politeness standards, that is, written in a well-mannered and nonoffensive way. Even if it criticizes the Member of Congress, it does so in a respectful way. For example: "you are going to have more of the same with HRC, and you are partly responsible. Very disappointed in all of you in DC" or "Fantastic article! I appreciate your understanding of the weaknesses of #medicaid, thanks for your leadership!"
2. *Uncivil*: an ill-mannered, disrespectful tweet that may contain offensive language. This includes threatening one's rights (freedom to speak, life preferences), assigning stereotypes or hate speech, name-calling ("weirdo," "traitor," "idiot"), aspersion ("liar," "traitor"), pejorative speak or vulgarity, sarcasm, ALL CAPS, and incendiary, obscene, and/or humiliating language. For example: "Just like the Democrat taliban party was up front with the AHCA. Hypocrites" or "Oh shut up David. You're a bore."

Following standard best practices for crowd-coding (Benoit et al., 2016), we used CrowdFlower's *quiz mode* option, which discards coders whose agreement with a small set of "gold" posts that we labeled manually is lower than 80%. This ensures that intercoder reliability is at a sufficiently high level: Average intercoder agreement was 89% in our final coded sample. In addition, we also provided coders with the following instruction:

> Note: we understand that civility can be subjective. Here we are looking not for your opinion, but rather what you think most people would respond in this situation. For example, some people may not find the word "weirdo" offensive, but generally it is considered impolite when it is used as an insult.

In our training set, 26% of tweets were labeled as uncivil.

Using our training data set, we then built supervised machine learning classifiers to predict the probability that each individual tweet is uncivil (as opposed to civil). As our initial attempts led to low performance, we increased the size of our training set by using synthetic labels for an additional set of 16,000 tweets using Google's Perspective API.[2] We trained a logistic regression with L1 regularization, also known as lasso (Friedman et al., 2001), using stemmed unigrams as features. We used fivefold cross-validation to identify the penalty parameter that maximizes in-sample performance.

By leaving out of the estimation a random sample of tweets that corresponds to 20% of our data set, and then assessing how well our classifiers perform on this "test set," we are able to evaluate the performance of our classifiers. We find that the overall accuracy is 90%, with precision and recall on the "civil" category being 92% and 95%, and precision and recall in the "uncivil" category being 73% and 61%.[3] Although the somewhat low recall on the uncivil category indicates that we may be underestimating the prevalence of incivility, note that this level of performance is typical in past studies that use machine learning to detect incivility (Davidson et al., 2017), given the inherent vagueness of this term and the heterogeneity of the language that can be used to express. However, overall these metrics clearly indicate that our method approximates the quality of human coding, which, as noted above, we estimated to be 89% based on the average intercoder agreement among our coders.

As an additional validation step for our measure of incivility, we also identified the top predictive n-grams for each category (see Table 1). As we expected, the classifier will predict as *uncivil* those tweets that contain insults and expletives, as well as words such as "idiot," "stupid," "moron," "disgrace," and "loser." In contrast, words that are related to civility do not follow any specific pattern, which is not surprising given that civility does not have any specific linguistic markers—we defined it based on the absence of offensive language. Beyond the performance metrics described above, these results confirm that our classifier is capturing our latent construct of interest.

Topic Modeling of Legislators' Tweets

We used a topic model to estimate the policy issue or political event that characterizes the content of each of the 155,540 tweets by Members of Congress published during our period of analysis. In particular, following Barberá et al. (2019), we fit a Latent Dirichlet Allocation model (Blei et al., 2003) with documents that aggregate all tweets by legislators for each party, day, and chamber. To identify the most appropriate number of topics (K), we ran separate models with K ranging from 20 to 120 at intervals of 10. For each value of K, we ran 10 different models and used cross-validation to determine how well the model fits the data on a 10% random sample of our data set that was not included in the estimation, computed using the perplexity metric (low values correspond to better fit). We also computed a metric of topic coherence that corresponds to the average within-topic cosine similarity in the word embeddings of the top 15 words divided by the average between-topic cosine distance on the same metric. This metric captures the extent to which each topic is internally consistent (because the embeddings of the top words are similar) but also separate from other topics

Table 1. Top Predictive Unigrams for Each Category.

Category	Predictive Unigrams
Uncivil	Idiot, stupid, moron, shit, hypocrit, ass, bullshit, fuck, asshol, pussi, suck, sccumbag, dick, crap, jerk, asshat, disgust, disgrac, radic, mulvaney, shitti, ugli, loser, scum, butt, garbag, racist, fascist, dumbass, min, #coward, blowhard, dumbest, witch, pedophil, lunat, cart, whore, arrog, no1, miser, ineffect, piti, fool, spineless, nee, blew, #li, abomin
Civil	pace, korean, commerci, convey, mitt, ilk, pari, #kremlinklan, ab, sto, atti, divers, furious, conway, arpaio, cohn, untru, huma, react, #msnbc, okay, ran, unfollow, toxic, oregon, cruel, refer, globalist, sea, regardless, latter, marco, pros, undeserv, #taxcutsfortherich, simpli, wil, #lockhimup, ff, repeat, tn, held, thank, ch, leadership, arizonan, there, band, access

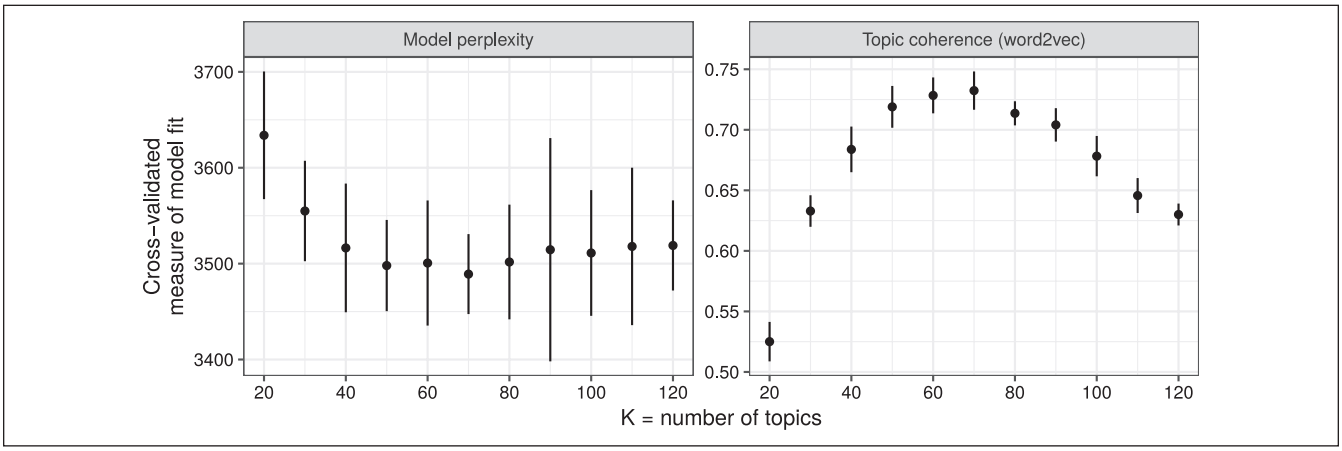

Figure 2. Choosing the right number of topics.

(because each topic occupies a separate location in the embedding space).

As we show in Figure 2, a value of $K = 70$ appears to yield the best results. From the 70 topics in the model, we manually selected the 45 topics that correspond to policy issues or political events and discarded the rest, which corresponded to nonpolitical topics or just general vocabulary words without a clear meaning. Figure 4 displays the top scoring words for each topic.

Metrics of Coordinated Activity

Our metric of *coordination* tries to capture whether uncivil tweeting is conducted by a small set of highly active users or instead by a broader sample of users who are on average not as active. We operationalize this concept by computing the Gini coefficient for inequality on the distribution of tweets by users (Barberá & Rivero, 2015). A Gini index close to 1 would mean that all uncivil tweeting is done by a single user and everyone else in our sample never tweets any uncivil tweet. A Gini index close to 0 would imply the opposite—all users are equally likely to send uncivil tweets targeting politicians. As Steinert-Threlkeld (2017) shows, when applied to Twitter data, this metric is able to capture coordination dynamics, and in particular whether a small set of users are leading some kind of organized action (which could either serve a grassroots campaign or serve a more pernicious purpose if it is a set of "trolls" trying to influence political conversations on Twitter) or whether in contrast the levels of involvement are approximately similar across all users.

Results

Figure 3 offers empirical evidence regarding the temporality of incivility on social media (Hypothesis 1). Here, we show that the level of vitriol addressed to Members of Congress on Twitter is relatively stable over time: Our machine learning model predicts that between 15 and 20% of tweets every day meet our definition of incivility.

However, the figure also reveals clear spikes of vitriol that are correlated over time, which suggests that specific events or policy debates are likely to spark cycles of higher incivility. We see one of those spikes in the month leading to the 2016 election. Based on an examination of a random sample of tweets during this period, it seems the increase is due to responses to anti-Clinton messages by Members of Congress such as Joe Walsh and Jason Chaffetz. Two subsequent spikes, in March and May 2017, can be clearly linked to the debate about the repeal of the Affordable Care Act and the new health care bill, which eventually failed. Another important spike, in August 2017, appears to correspond to reactions to the White supremacist rally in Charlottesville and the

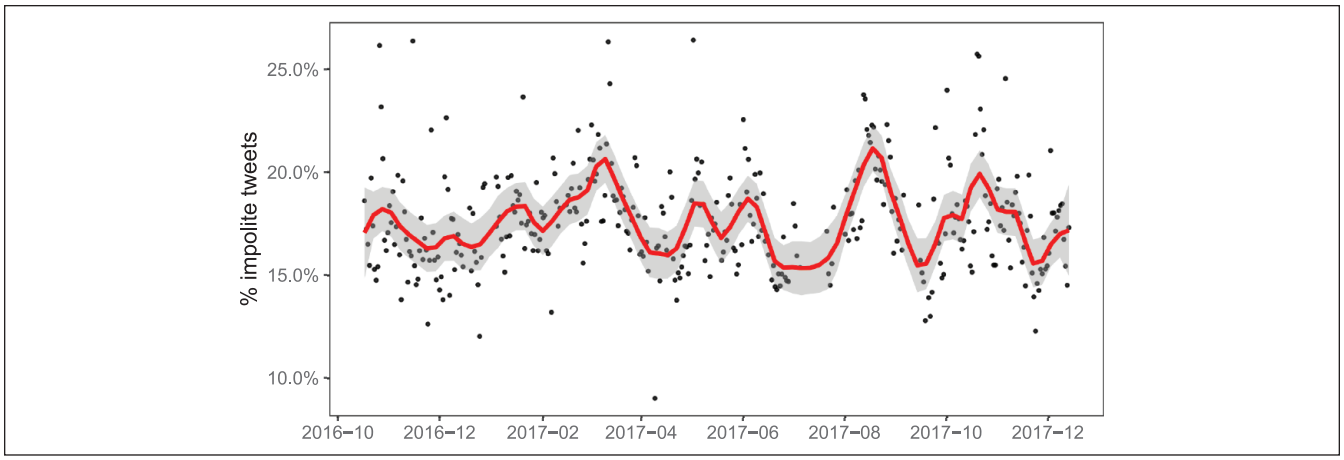

Figure 3. Incivility in tweets mentioning U.S. legislators over time.

subsequent reactions to it by legislators. We observe one last spike in October 2017, which overlaps with the revelation by Florida congresswoman Frederica S. Wilson that President Trump disrespected the family of a U.S. soldier killed in an attack in Niger.

In conclusion, we find mixed evidence in support of Hypothesis 1: Although incivility is always prevalent on Twitter, we also find highly communicative events that lead to significant increases in vitriol and attack targeting politicians.

Our second hypothesis (Hypothesis 2) explores variation at the meso level, in response to the content of the messages that politicians send. Figure 4 summarizes the analysis we conducted to test this hypothesis. For each of the 45 political topics we identified in legislators' tweets, we fit a linear regression model where the dependent variable is the proportion of uncivil responses that legislators received to each of their tweets, and the independent variable corresponds to the probability that a given tweet falls into the topic in consideration. In our models, we also control for the chamber and party of the politician as well as for the total number of replies that each tweet received. Figure 4 reports only the coefficient for the topic effect from each of the 45 separate regressions. In other words, the coefficients here can be interpreted as the predicted increase in the proportion of replies to a given tweet that are uncivil if that tweet is completely about a particular topic.

This analysis reveals large differences in incivility across topics. Contrary to our expectations, when we manually divide the topics into policy issues (health care, immigration, gun control, taxes, etc.) and political events and scandals (Trump's inauguration, the Clinton email scandal, Supreme Court nominations, hurricanes, etc.), we find that political topics lead to a somewhat larger increase in incivility than policy topics—on average, the coefficient for the former is .16 but .06 for the latter. These results are also consistent with our earlier finding regarding spikes on incivility: Here, we also find that tweets about the Charlottesville White nationalist rally sparked the most uncivil responses.

Turning to the micro level, we were also interested in examining whether uncivil tweeting is carried out by a small minority of users (Hypothesis 3). As it is usually the case on social media, we do find that this holds here too: The 1% (10%) most active users are responsible for 20% (56%) of all uncivil tweets. The Gini coefficient for the entire distribution is .67. However, this level of coordination is actually smaller than if we look at the full data set of tweets mentioning politicians, where the top 1% (10%) most active users produce 29% (69%) of all tweets and the Gini coefficient is .77.

Figure 5 visualizes the full distribution of users and tweets they produce using a Lorenz curve. Overall, this suggests that even if it is indeed the case that a small number of users are responsible for most of the content, this level of concentration is not higher than for other types of tweets. In fact, up to 38% of all the unique users who tweeted at least once retweeting or mentioning the name of a politician (648,690 out of 1,686,540 users) sent at least one tweet that our classifier predicted was uncivil. In other words, consistent with previous findings by McLoughlin and Ward (2017), we do not find that organized trolls or "serial transgressors" are responsible for most of the abuse; on the contrary, uncivil behavior seems to be common among many frustrated citizens who take to Twitter to angrily criticize and insult politicians.

However, that does not mean that trolling campaigns do not exist. When we measure coordination at the day level using the Gini coefficient for the distribution of users and tweets sent each day, shown in Figure 6, we do find clear variation over time. The correlation between this time series and the proportion of uncivil tweets at the day level (in Figure 3) is .26, which suggests that part of the spikes we described earlier could be due to an increase in the level of coordination among users. And indeed, some of the days with the highest values on our coordination metric coincide

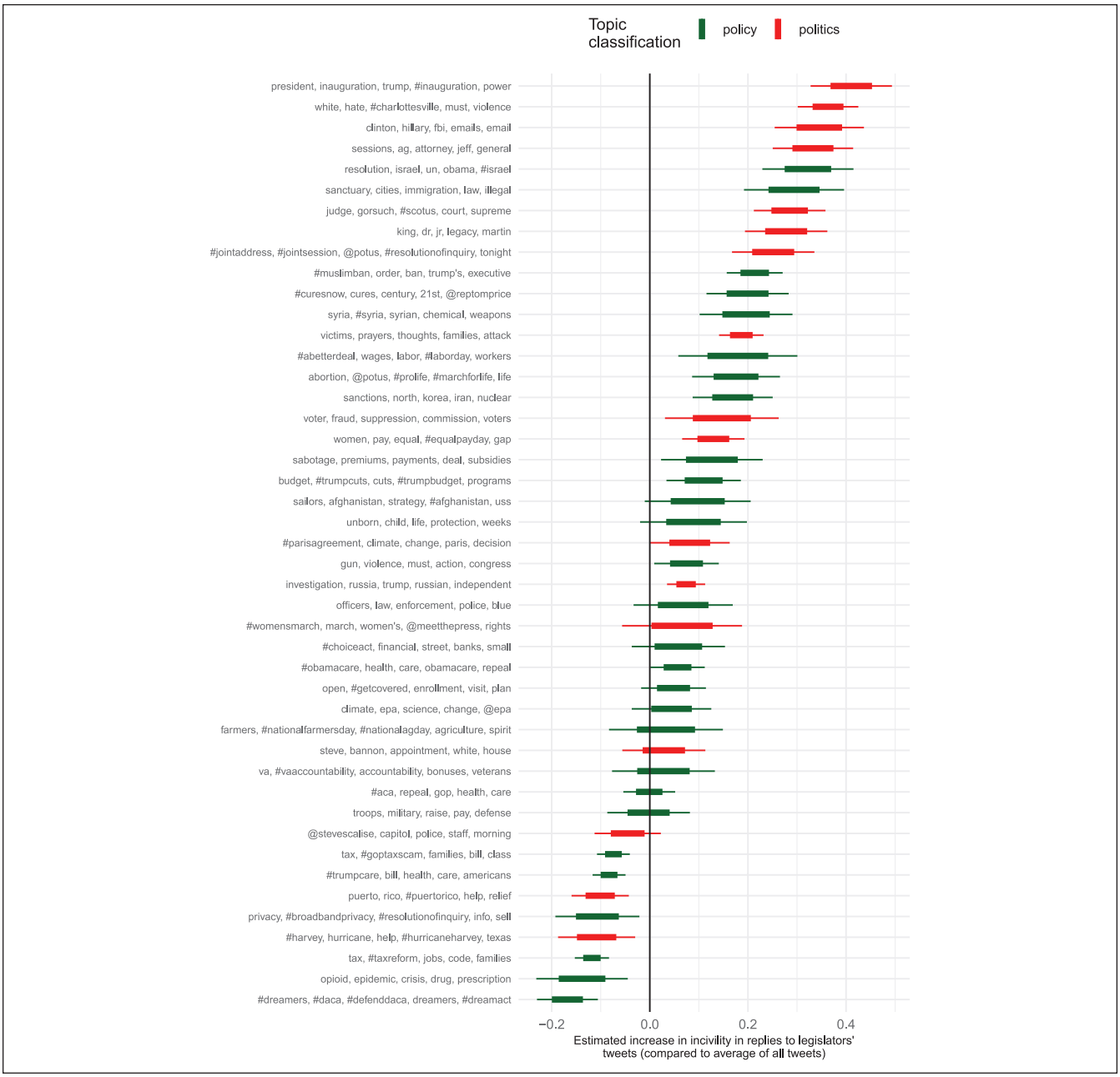

Figure 4. Incivility in replies to legislators' tweets, depending on topic of root tweet.

with the same set of external events we identified as motivating spikes in incivility: for example, the attempted repeal of Obamacare in early March and early May 2017 or the reactions to the Charlottesville White supremacist rally in August 2017. This means that even if incivility is a constant presence on political tweets, we do find that its prevalence appears to be due to coordinated campaigns taking place at specific periods of time.

To further explore the motives for these coordinated campaigns, we now turn our unit of analysis from the individual tweet to the target of the tweet and, in particular, the legislator that is mentioned on it. We will try to identify which characteristics predict who is most likely to be receiving uncivil tweets. Table 2 reports the results of this analysis. Here, we show results of multivariate regression models where the dependent variable is the percentage of uncivil tweets addressed to each of the Members of U.S. Congress in our sample (Models 1 and 2) or the level of coordination in those tweets, measured as the Gini coefficient (multiplied by 100 to facilitate interpretation) for the distribution of users and uncivil tweets sent (Models 3 and 4). The main independent variables measure different characteristics of each

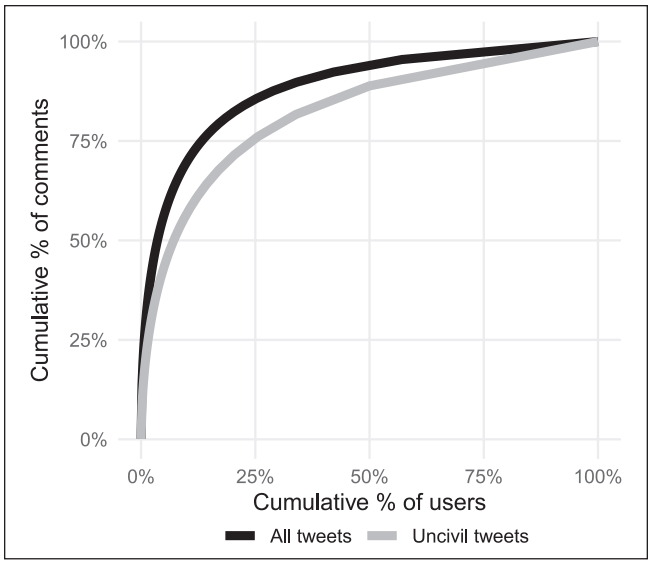

Figure 5. Lorenz curves: Inequality in production of uncivil tweets versus all tweets.

politician: gender, type (senator or representative), party affiliation, and extremism, which we measure as the absolute value of DW-NOMINATE (a metric of ideology estimated using roll-call votes in Congress; see Lewis et al., 2018). We also control for the (logged) number of tweets mentioning each of these politicians in our sample and, in our second set of models, for the average incivility in the tweets addressed to each politician (the dependent variable in the first two models).

We do not find clear evidence supporting our hypothesis stating that female politicians were more likely than men to receive uncivil comments (Hypothesis 4). The sign of the coefficient for gender is in the expected direction in Models 1 and 3, but it is not statistically significant in any of the models.

In contrast, chamber and ideological position appear to have large effects on the overall level of incivility received. Members of the House of Representatives who adopt a more extreme ideological position are much more likely to receive uncivil comments. To illustrate the substantive magnitude of this effect, note that the predicted difference between Senators with an extremism score of 0 and Representatives with an extremism score of 1 is 7.3 percentage points (12.3 vs. 4.8 percentage points). This is a substantively large effect that represents approximately one third of the average proportion of incivility at the day level. When it comes to our regression models for coordination, we find that only party identification is a significant predictor: Republican legislators appear to be more likely to be the subject of coordinated attacks on Twitter.

Discussion

Online incivility and harassment in political communication are a growing concern among politicians, journalists, and academics. This article provides a descriptive account of how incivility evolves over time, who is the target of uncivil comments on Twitter, and whether being a target is correlated with individual-level characteristics. We relied on a combination of supervised and unsupervised machine learning methods to examine these questions.

Our empirical findings illuminate past unresolved debates in the literature and also challenge the conventional wisdom within this field. We find that the prevalence of incivility in social media conversations that involve politicians is relatively constant over time, although we also find spikes that correspond to external events. These spikes tend to overlap with the emergence of controversial issues or events that spark outrage and lead ordinary citizens to address vitriolic messages to their representatives. This is in line with previous studies stressing the role of political expression on social media in the influence of temporal dynamics (Jungherr,

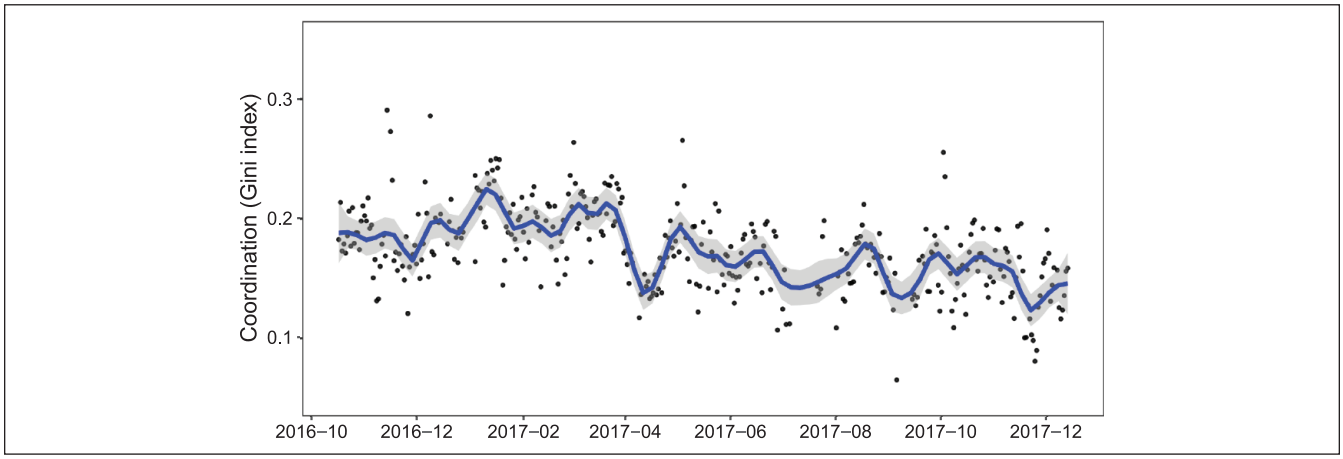

Figure 6. Coordination of incivility in tweets mentioning U.S. legislators over time.

Table 2. OLS Regressions of Incivility and Coordination on Legislators' Characteristics.

Covariate	Model 1	Model 2	Model 3	Model 4
Intercept	7.39* (0.81)	–4.87* (1.05)	16.74* (1.71)	–8.57* (2.35)
Gender = female	0.35 (0.63)	–0.04 (0.52)	0.76 (1.33)	–0.28 (1.12)
Type = senator	–2.50* (0.62)	–3.85* (0.52)	0.31 (1.31)	–1.88 (1.18)
Party = independent	–2.24 (3.89)	–3.08 (3.24)	–5.96 (8.18)	–7.55 (6.92)
Party = Republican	0.49 (0.53)	0.17 (0.44)	5.77* (1.12)	5.00* (0.95)
Extremism (0–1)	6.49* (1.75)	3.48* (1.47)	0.62 (3.69)	–6.75* (3.18)
Log (no. of mentions)		1.70* (0.11)		3.41* (0.30)
Average incivility				0.10 (0.09)
N	518	518	512	512
R^2	.07	.36	.06	.33
Adjusted R^2	.06	.35	.05	.32
Resid. *SD*	5.42	4.51	11.40	9.64

Note. Standard errors in parentheses. DV in Models 1 and 2 is percentage of uncivil tweets targeting a legislator. DV in Models 3 and 4 is Gini coefficient (multiplied by 100) of distribution of uncivil tweets sent by user. DV = dependent variable; OLS = ordinary least squares.
*Significant at $p < .05$.

2014). Contrary to the conventional wisdom in media accounts and the academic literature, our analysis suggests that uncivil tweets can be sent by any user and not only "professional trolls" who engage in coordinated action (although we also find evidence of such attacks in our analysis). This finding has important implications when it comes to addressing the issue of coordinated attacks and bot inferences in political debates, as it makes it unclear whether measures toward reducing such threats can do much in addressing the issue of toxic language. Finally, we do find variation regarding which politicians are more likely to be victims of attacks, although gender does not have a statistically significant effect.

Although our analysis offers what we believe is the most systematic descriptive evidence of incivility on Twitter, we also caution readers that it may not be without limitations. First, our finding regarding the lack of differences across gender groups might be partially due to measurement error—probably because we have more male politicians than female politicians in our training set, the performance of our classifier of uncivil tweets is somewhat lower when the tweet is addressed to female politicians (90% vs. 88% accuracy; 62% vs. 54% recall on the uncivil category). Performance may also be lower if the insults used to attack female politicians are qualitatively different than those used to attack male politicians, which is likely to be the case. We address this limitation in a follow-up project where we code an equal number of tweets targeting each gender group (Theocharis et al., 2018).

And second, we caution against generalizing our findings to other cases or periods. For example, it is possible that the differences across parties are due to the current political context in the United States, with a Republican president and a Republican majority in both House and Senate. We would need to replicate our analysis with past or future data to determine whether Democrats in Congress may be subject to higher levels of incivility if they held power. Similarly, it is hard to tell how these findings may be different in other countries. Theocharis et al. (2016) found significant differences in the level of incivility on Twitter across four different European countries. It is likely that some of the patterns identified here, such as the variation in the level of coordination or regarding who is the target of harassment, are also different in other contexts.

Despite these caveats, we claim that our analysis offers an important account of the dynamics of incivility toward politicians on social media. As citizens increasingly rely on these platforms to consume political news and to engage in conversations about politics, we believe that understanding the determinants of the high levels of toxicity in these sites is a necessary first step toward design changes that may turn social media platforms into public spaces for deliberation where everyone feels free to intervene and share their opinion.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Yannis Theocharis https://orcid.org/0000-0001-7209-9669
Pablo Barberá https://orcid.org/0000-0002-9063-4829

Notes

1. We use the *follow* parameter in the *filter* endpoint and query the legislators' user IDs directly instead of their Twitter handles to minimize data loss.

2. More in detail, we ran all 20,000 tweets (4,000 in the original training set and 16,000 additional tweets) through Google's Perspective API (application programming interface), which generates predictions for a variety of negative speech categories, such as "toxic," "hate speech," "attacks," and "profanity". Then, we fit a classifier with the 4,000 original labels using Google's Perspective API predictions as features. Not surprisingly, given the similarity in the categories being predicted, we obtained high accuracy (83% accuracy, 0.80 AUC [area under the receiver operating characteristic curve]). We then predicted the labels for all the remaining 16,000 tweets, which we will use to train the classifier to predict incivility in our entire corpus.
3. Accuracy is the percentage of tweets correctly classified. Precision is the percentage of tweets predicted to be in a given category that are correctly classified. Recall is the percentage of tweets in a given category (according to human annotators) that are correctly classified.

References

Amnesty International. (2018). *Toxic Twitter: A toxic place for women*.

Anderson, A. A., Brossard, D., Scheufele, D., Xenos, M., & Ladwig, P. (2013). The "nasty effect": Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, *19*(3), 373–387.

Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, *113*(4), 1–19.

Barberá, P., Jost, J., Nagler, J., Tucker, J., & Bonneaur, R. (2015). Tweeting from left to right is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542.

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, *33*(6), 712–729.

Barberá, P., & Zeitzoff, T. (2017). The new public address system: Why do world leaders adopt social media? *International Studies Quarterly*, *62*(1), 121–130.

BBC. (2016, November 23). Calgary politician Sandra Jansen denounces online hate.

BBC. (2017, March 11). Owen Jones quits social media after threats of "torture and murder."

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, January). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bode, L., & Dalrymple, K. E. (2016). Politics in 140 characters or less: Campaign communication, network interaction, and political participation on Twitter. *Journal of Political Marketing*, *15*(4), 311–332. https://doi.org/10.1080/15377857.2014.959686

Borah, P. (2012). Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research*, *46*(6), 809–882.

Brooks, D., & Geer, J. (2007). Beyond negativity: The effects of incivility on the electorate. *American Journal of Political Science*, *51*(1), 1–16.

Capella, J. N., & Jamieson, K. H. (1997). *Spiral of cynicism: The press and the public good*. Oxford University Press.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*, 658–679.

Coleman, S. (2005). New mediation and direct representation: Reconceptualizing representation in the digital age. *New Media & Society*, *7*(2), 177–198.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language In *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17* (pp. 512–515). Association for the Advancement of Artificial Intelligence.

Elving, R. (1994). *Brighter lights, wider windows: Presenting Congress in the 1990s* [Technical report]. American Enterprise Institute, the Brookings Institution.

Enli, G., & Skogerbø, E. (2013). Personalized campaigns in party-centred politics. *Information, Communication & Society*, *16*(5), 757–774.

Every-Palmer, S., Barry-Walsh, J., & Pathe, M. (2015). Harassment, stalking, threats and attacks targeting New Zealand politicians: A mental health issue. *Australian and New Zealand Journal of Psychiatry*, *49*, 634–641.

Fazekas, Z., Popa, S. A., Schmitt, H., & Barberá, P. (2018). *Elite-public interaction on Twitter: EU issue expansion in the campaign*. SGEU Standing Group on the European Union.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer Series in Statistics.

Frijda, N. (1988). The laws of emotion. *American Psychologist*, *43*, 349–358.

Funk, C. A. (2001). Process performance: Public reaction to legislative policy debate. In J. R. Hibbing & E. Theiss-Morse (Eds.), *What is it about government that Americans dislike?* (pp. 193–204) Cambridge University Press.

Gervais, B. (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.

Graham, T., Broersma, M., Hazelhoff, K., & van't Haar, G. (2013). Between broadcasting political messages and interacting with voters. *Information, Communication & Society*, *16*(5), 692–716. https://doi.org/10.1080/1369118X.2013.785581

Graham, T., Jackson, D., & Broersma, M. (2014). New platform, old habits? Candidates' use of Twitter during the 2010 British and Dutch general election campaigns. *New Media & Society*, *18*(5), 765–783.

The Guardian. (2016, May 31). Labour MP says she may leave Twitter over trolls' rape abuse. https://www.theguardian.com/technology/2016/may/31/labour-mp-jess-phillips-says-she-may-leave-twitter-over-trolls-abuse

Gulati, J., & Williams, C. B. (2010). Congressional candidates' use of YouTube in 2008: Its frequency and rationale. *Journal of Information Technology & Politics*, *7*(2), 93–109.

Gulati, J., & Williams, C. B. (2013). Social media and campaign 2012: Developments and trends for Facebook adoption. *Social Science Computer Review*, *31*, 577–588.

Halliday, J. (2012, August 2). *Helen Skelton quits Twitter after abuse from trolls*. https://www.theguardian.com/technology/2012/aug/02/celebrities-quit-twitter-abuse

Hay, C. (2007). *Why we hate politics*. Polity Press. ISBN 9780745630991 (pbk) 0745630995 (pbk) 9780745630984 (hbk) 0745630987 (hbk).

Hayes, R. (2008, November 21–23). *Providing what they want and need on their own turf: Social networking, the web, and young voters* [Paper presentation]. Association Annual Conference, San Diego, CA.

Heiss, R., Schmuck, D., & Matthes, J. (2018). What drives interaction in political actors' Facebook posts? Profile and content predictors of user engagement and political actors' reactions. *Information, Communication & Society*, *22*(10), 1497–1513. https://doi.org/10.1080/1369118X.2018.1445273

Henson, B., Reyns, B. W., & Fisher, B. S. (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*, *29*(4), 475–449.

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Temple University Press.

Hinduja, S., & Patchin, J. W. (2007). Offline consequences of online victimization. *School Violence and Delinquency*, *6*(3), 89–112.

Inter-Parliamentary Union. (2016). *Sexism, harassment and violence against women parliamentarians* [Technical report].

Jackson, N., & Lilleker, D. (2011). Microblogging, constituency service and impression management: UK MPs and the use of Twitter. *The Journal of Legislative Studies*, *17*(1), 86–105.

James, D. V., Farnham, F., Sukhwal, S., Jones, K., Carlisle, J., & Henley, S. (2016). Aggressive/Intrusive behaviours, harassment and stalking of members of the United Kingdom parliament: A prevalence study and cross-national comparison. *The Journal of Forensic Psychiatry & Psychology*, *27*(2), 177–197.

Jamieson, K. H., Volinsky, A., Weitz, I., & Kenski, K. (2017). The political uses and abuses of civility and incivility. In K. Kenski & K. H. Jamieson (Eds.), *The Oxford handbook of political communication* (pp. 205–218). Oxford University Press.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication*, *64*(2), 239–259. https://doi.org/10.1111/jcom.12087

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, *13*(1), 72–91.

Kamps, H. J. (2016). *Solving Twitter's abuse problem*. https://medium.com/@Haje/solving-twitter-s-abuse-problem-3f1f8ac1a0d2

Karlsen, R., & Skogerbø, E. (2015). Candidate campaigning in parliamentary systems. *Party Politics*, *21*(3), 428–439.

Kittilson, M. C., & Fridkin, K. (2008). Gender, candidate portrayals and election campaigns: A comparative perspective. *Politics & Gender*, *4*(3), 371–392. https://doi.org/10.1017/S1743923X08000330

Larsson, A. (2015). The EU parliament on Twitter—Assessing the permanent online practices of parliamentarians. *Journal of Information Technology & Politics*, *12*(2), 149–166.

Lee, E. J., & Shin, S. (2012). Are they talking to me? Cognitive and affective effects of interactivity in politicians' Twitter communication. *Cyberpsychology & Behavior, and Social Networking*, *15*(10), 515–520.

Lewis, J. B., Poole, K., Rosenthal, H., Boche, A., Rudkin, A., & Sonnet, L. (2018). Voteview: Congressional roll-call votes database. https://voteview.com

Lühiste, M., & Banducci, S. (2016). Invisible women? Comparing candidates' news coverage in Europe. *Politics & Gender*, *12*(2), 223–253. https://doi.org/10.1017/S1743923X16000106

Lyons, B., & Veenstra, A. (2016). How (not) to talk on Twitter: Effects of politicians' tweets on the whole Twitter environment. *Cyberpsychology, Behavior, and Social Networking*, *19*(1), 8–15.

Mantilla, K. (2015). *Gendertrolling: How misogyny went viral*. Praeger.

Marwick, A. E., & Lewis, R. (2017). *Media manipulation and disinformation online* [Technical report]. Data & Society.

Massaro, T. M., & Stryker, R. (2012). Freedom of speech, liberal democracy, and emerging evidence on civility and effective democratic engagement. *Arizona Law Review*, *54*(2), 375–411.

McLoughlin, L., & Ward, S. (2017, April 25–29). *Turds, traitors and tossers: The abuse of UK MPs via Twitter* [Paper presentation]. European Consortium of Political Research Joint Sessions, Nottingham.

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication*, *11*, 3182–3202.

Munger, K. (2016). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649.

Munger, K., Egan, P., Nagler, J., Ronan, J., & Tucker, J. A. (2016). Learning (and unlearning) from the media and political parties: Evidence from the 2015 UK election. (Unpublished manuscript). New York University.

Mutz, D. C. (2015). *In-your-face politics: The consequences of uncivil media*. Princeton University Press.

Mutz, D. C., & Reeves, B. (2005). The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review*, *99*(1), 1–15.

Nulty, P., Theocharis, Y., Popa, S. A., Parnet, O., & Benoit, K. (2016). Social media and political communication in the 2014 elections to the European Parliament. *Electoral Studies*, *44*, 429–444.

Oluo, I. (2016, July 19). Leslie Jones' Twitter abuse is a deliberate campaign of hate. *The Guardian*. https://www.theguardian.com/commentisfree/2016/jul/19/leslie-jones-twitter-abuse-deliberate-campaign-hate

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society*, *20*(9), 3400–3419. https://doi.org/10.1177/1461444817749516

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283.

Pelled, A., Lukito, J., Boehm, F., Yang, J., & Shah, D. (2018). "Little Marco," "Lyin' Ted, "Crooked Hillary," and the "Biased" media: How Trump used Twitter to attack and organize. In N. J. Stroud & S. C. McGregor (Eds.), *Digital discussions: How big data informs political communications* (pp. 176–196). Routledge.

Pham, S. (2017, February 7). Twitter tries new measures in crackdown on harassment. *CNN Business*. https://money.cnn.com/2017/02/07/technology/twitter-combat-harassment-features/

Phillips, W. (2016). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.

Popa, S. A., Fazekas, Z., Braun, D., & Leidecker-Sandmann, M. M. (2019). Informing the public: How party communication builds opportunity structures. *Political Communication*, 1–21. https://doi.org/10.1080/10584609.2019.1666942

Rheault, L., Rayment, E., & Musulan, A. (2019). Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics*, *6*(1), 1–7. https://doi.org/10.1177/2053168018816228

Rossini, P. G. C. (2019). Disentangling uncivil and intolerant discourse. In R. G. Boatright, T. J. Shaffer, S. Sobieraj, & D. G. Young (Eds.), *A crisis of civility? Contemporary research on civility, incivility and political discourse*. Routledge.

Saner, E. (2016, June 18). Vile online abuse against female MPs "needs to be challenged now." *The Guardian*. https://www.theguardian.com/technology/2016/jun/18/vile-online-abuse-against-women-mps-needs-to-be-challenged-now

Sapiro, V. (1999). *Considering political civility historically: A case study of the United States*. Boston University.

Sigelman, L., & Bullock, D. (1991). Candidates, issues, horse races, and hoopla: Presidential campaign coverage, 1888–1988. *American Politics Research*, *19*(1), 5–32.

Sobieraj, S., & Berry, J. (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication*, *28*(1), 19–41.

Steinert-Threlkeld, Z. C. (2017). Spontaneous collective action: Peripheral mobilization during the Arab Spring. *American Political Science Review*, *111*(2), 379–403.

Stromer-Galley, J. (2000). Online interaction and why candidates avoid it. *Journal of Communication*, *50*(4), 111–132.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, *7*(3), 321–326.

Sydnor, E. (2019). *Disrespectful democracy: The psychology of political incivility*. Columbia University Press.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil Twitter use when interacting with party candidates. *Journal of Communication*, *66*(6), 1007–1031.

Theocharis, Y., Luhiste, M., Fazekas, Z., Popa, S. A., & Barbera, P. (2018, May 24-28). *When does abuse and harassment marginalize female political voices on social media?* Annual Meeting of the International Communication Association, Prague, Czech Republic.

Torcal, M., & Montero, J. R. (2006). *Political disaffection in contemporary democracies: Social capital, institutions, and politics*. Routledge.

Tromble, R. (2018). Thanks for (actually) responding! How citizen demand shapes politicians' interactive practices on Twitter. *New Media & Society*, *20*(2), 676–697. https://doi.org/10.1177/1461444816669158

Uslaner, E. (1993). *The decline of comity in Congress*. University of Michigan Press.

Ward, S., & Lusoli, W. (2005). "From weird to wired": MPs, the Internet and representative politics in the UK. *Journal of Legislative Studies*, *11*(1), 57–81.

Williams, C. B., & Gulati, J. (2010, September 2-5). *Communicating with constituents in 140 characters or less: Twitter and the diffusion of technology innovation in the United States Congress*. Annual Meeting of the American Political Science Association, Washington DC.

Williamson, A. (2010). *Digital citizens and democratic participation: An analysis of how citizens participate online and connect with MPs and parliament* [Technical report]. Hansard Society.