It is commonplace to find abusive language directed toward women and racial or sexual minorities online. These types of messages often reverberate through networks of interacting users (Felmlee, Inara Rodis, and Francisco 2018), with negative tweets reposted more rapidly and frequently than positive or neutral messages (Tsugawa and Ohsaki 2017). In response to public outcry, Twitter recently embarked on widely publicized policy changes to clarify existing standards and facilitate more reporting and removal of abusive content. We examine the effect of one set of major policy rule changes on negative messages through a sentiment analysis of tweets sent before and after the policy shift.

This moment provides a timely opportunity to examine how the content of derogatory messages may be influenced by Twitter’s attempt to exert greater control over its users. Here we build on this opportunity to test several social scientific arguments regarding the influence of organizational efforts to sanction negative behavior. Previous research suggests that such top-down attempts at control can be effective if the sanctioning authority is perceived as legitimate (Baldassarri and Grossman 2011). However, other evidence suggests that targeted bullies and trolls may instead respond with even more aggressive behavior or take covert measures to evade detection (Flores 2017).

**Online Harassment on Social Media**

Online harassment constitutes a routine occurrence in society, where online harassment refers to unwanted, problematic digital contact (Lenhart et al. 2016), and that arises in various forms with numerous labels, such as cyberbullying, hate speech, and online aggression. According to nearly two thirds of Americans (62 percent), online harassment poses a major social problem (Pew Research Center 2017), with close to half (47 percent) experiencing online harassment personally (Lenhart et al. 2016). These types of problematic online interactions can be harmful, inflicting emotional and psychological damage on their victims (e.g., Juvonen and...
The threat of possible Internet abuse also can lead individuals, especially former victims, to self-censor and suppress their online presence (Lenhart et al. 2016), suggesting that online harassment curtails communication, particularly for those most vulnerable to attack.

Studies of cyberbullying and harassment tend to take an individual-level focus and could benefit from more of a sociological lens, according to scholars (Ey, Taddeo, and Spears 2015). Recent research points to two social, group processes that contribute to cyber and peer aggression, namely, the enforcement of social norms and the enhancement of social status (Faris and Felmlee 2011, 2014; Felmlee and Faris 2015). Recent research points to two social, group processes that contribute to cyber and peer aggression, namely, the enforcement of social norms and the enhancement of social status (Faris and Felmlee 2011, 2014; Felmlee and Faris 2015). For organizations to be effective in limiting harassment and abuse on their social media platforms, such findings suggest that these organizations will need to shift norms to encourage greater online civility and/or make harassment less socially rewarding by minimizing its virtual footprint. Indeed, removal of offending messages could curtail the ability of hostile users to expand their online followings via likes and content sharing and thereby diminish one type of motivation for disseminating disparaging content.

Several previous studies (e.g., Chatzakou et al. 2017; Felmlee et al. 2018; Felmlee, Inara Rodis, and Zhang 2020; Xu et al. 2012) examined online harassment on Twitter, where racist, sexist, and homophobic tweets remain readily accessible to the general public at any time of day. Prior research finds that it takes an average of 24 seconds to 1.5 minutes to locate the first of thousands of derogatory, aggressive tweets, on the basis of a sample of searches for assorted defamatory slurs and insults (e.g., n**ger, who*e, fa**ot) (Sterner and Felmlee 2017). Additional research on sexist tweets identified more than 419,000 tweets per day that included at least one of four sexist slur words (bi*ch, c**t, s**t, and who*e) during a one-week collection period (Felmlee et al. 2020). Such negative messages often can be retweeted or “liked” by followers, thereby creating networks of cyberbullying that spread far beyond the original perpetrator and target (Felmlee et al. 2018). Studying tweets over two years, researchers found that people posted more than 9.76 million tweets referencing bullying on Twitter (Bellmore et al. 2015), which averages to more than 130,000 discussions of bullying reverberating online each day. According to additional research, moreover, approximately 10,000 racial slurs occur in tweets on a daily basis (Bartlett et al. 2014). Furthermore, news accounts repeatedly highlight the outcry surrounding damaging messages on Twitter. In one instance, a number of women participated in a Twitter boycott (Women’s Boycott) triggered by the suspension of the Twitter account of an actress, Rose McGowan, who posted allegations of sexual abuse by Hollywood producer Harvey Weinstein.

Following the Women’s Boycott and repeated claims of online harassment and abuse, the CEO of Twitter pledged to introduce new “anti-abuse” rules that would be “more strictly enforced” over the coming weeks (October 14, 2017). The new rules would apply to unwanted sexual advances, nonconsensual nudity, hate symbols, violent groups, and tweets that glorify violence. In addition to updating and clarifying rules, Twitter announced changes to the content reporting and removal processes to be implemented, and these policy adjustments were widely publicized in media outlets and in a “Safety Calendar” posted on Twitter’s blog. These new rules were announced as going into effect on December 18, 2017.

One notable feature of Twitter’s attempts to affect its users’ behavior is that the organization did not outright censor the use of specific slurs or derogatory terms (e.g., bi*ch, n**ger, fa**ot). This likely owes to the perception that there is “gray area” in the use of such words, particularly depending on who specifically is using them. Sometimes members of the targeted group themselves (e.g., women, African Americans, or LGBTQ people) use these terms in a process that the feminist writer Judith Butler (1990) called “resignification.” In other cases, slurs allow an individual to express a positive message while simultaneously using the slur as a means of self-protection or boundary drawing (Pascoe and Diefendorf 2019). Instead, Twitter’s attempts to reduce offensive content rely on a combination of top-down organizational signals that abusive behavior is considered non-normative and unacceptable, as well as bottom-up peer reporting or “flagging” of offensive material. Without outright banning certain offending terms, Twitter thus attempts to regulate how such terms are used on the platform.

Organizational Policies and Individual Responses

In the present study we ask: Did Twitter’s rule changes influence patterns of abusive language on the platform? Previous work in sociology, political science, and economics suggests that such changes are potentially efficacious but that this effectiveness depends on characteristics of the environment into which the rule is introduced. Institutionalist models of social control emphasize the importance of the sanctioning authority’s perceived legitimacy in dictating users’ behavior (e.g., Baldassarri and Grossman 2011; Fehr and Gachter 2002; Henrich et al. 2006). If users perceive Twitter as a legitimate referee or institutional arbiter empowered to set the “rules of the game” (e.g., North 1990), then a top-down rule change can effectively signal a shift in norms of acceptable behavior. In a basic sense, the effect of such legitimate authority is analogous to the authority a state exercises in crime and punishment. Just as the state punishes lawbreaking with incarceration (removal from free society), the legitimate Twitter administrators might punish rule breaking with (in some cases) removal from the platform (Tyler et al. 2018). Although some especially belligerent users may still engage in aggressive behavior, others will comply so as not to risk their continued access to the platform.
However, previous theory and evidence suggests that the dynamics of rule and norm following on Twitter and other social media platforms are not likely to be so straightforward. First, to sanction abusive behavior in the first place, Twitter users must be aware of it. To this end, the organization needs to rely on individual users’ willingness to report abuses (though natural language processing advances may soften this requirement; e.g., Davidson et al. 2017). Thus, the top-down rule change will achieve maximum effect only if supported by a bottom-up movement of Twitter users mobilized to maintain social norms and report aggressive behavior to authorities. The presence or absence of such mobilization may depend on both the value users attach to norms of decent and nonabusive behavior and on the network externalities associated with violations of these norms (DellaPosta, Nee, and Opper 2017). These network externalities depend in turn on the audience that given Twitter users believe themselves to be facing when they tweet (Marwick and boyd 2011). If Twitter “bullies” are embedded in dense networks of fellow “bullies,” they may perceive social rewards in the continued negative targeting of others. However, if a shift in perceived rules and norms on the platform leads perpetrators to have greater fear of exposure and sanction, they might respond by engaging in self-censorship that would decrease the overall flow of negative messaging on the platform.

A recent study by Flores (2017) suggests that increased aggression may be an equally plausible outcome of Twitter’s attempts to crack down on negative messages. Examining Twitter messages before and after Arizona’s passage of an anti–illegal immigration measure, SB 1070, Flores found a mobilization of anti-immigrant messages attributable to the activation of both new and previous Twitter users. Although Flores’s study shows platform users’ mobilizing in response to an external societal signal that supported their already-held views, it is not difficult to imagine the reverse—aggressive or abusive “trolls” ratcheting up “psychological reactance” (Brehm 1966) against the perceived heavy-handedness of Twitter authorities. If other users respond to the rule changes with vocal support, the targeted trolls may then develop an even stronger sense of belonging to an oppositional identity movement (e.g., Bearman and Brückner 2001). Tweeting abusive language will then become a salient symbolic act in and of itself. Bullies may send abusive tweets to signal their social identity and opposition to the perceived oppression of political correctness (DellaPosta, Shi, and Macy 2015; Gusfield 1966; Smaldino et al. 2017).

Reactance dynamics of this kind are widely observed in studies of employee responses to top-down change in organizational contexts. When an organization’s administration institutes policies that threaten the perceived freedom and autonomy of members, these members often respond with deliberative noncompliance and resistance to reassert their freedom to behave as they wish (e.g., Lowry and Moody 2015; Nesterkin 2013). It is plausible to think that users of social media platforms would respond similarly to organizational attempts to control their behavior. Users—and even many of the people who control social media platforms (Massanari 2017)—often see social media sites such as Twitter, Facebook, and Reddit as occupying the role of a neutral platform where all views are equally valid to share and where top-down censorship would violate a right to “free speech.” Policies that restrict a user’s autonomy to post negative content go against such a conception of the platform as impartial and are thus likely to provoke reactance among at least a vocal subset of users.

Study Design

In light of these competing intuitions, the purpose of our research is to examine the degree to which Twitter’s rule changes influenced the content and network spread of aggressive messages, focusing on content related to gender and race. Did the organization’s policies alter the level of hostility in messages? To address this question, we begin by locating tweets that include a gendered slur, bi!ch, during the month prior to and following the effective date of the most widely publicized of Twitter’s recent rule implementations (December 18, 2017). Previous research finds that bi!ch is a particularly common slur used to demean a feminine target on Twitter (Felmlee et al. 2020), and it is one of the four most prevalent curse words on the platform (Wang et al. 2014).

In response to calls for further attention to online bullying of racial minorities (Zych, Ortega-Ruiz, and Del Ray 2015), next we examine tweets during that period that contain a racist slur, n!/ger. We chose that particular offensive term because it is one of the slurs used most frequently to demean individuals on the basis of race (e.g., Chaudhry 2015; Felmlee et al. 2018; Wang et al. 2014). Although bi!ch and n!/ger are both top-curse words used in online messages, our choice of these two terms does not imply that they are equivalent. Quite the opposite, we believed that examining these two terms would lend our analysis greater robustness by showing how Twitter’s organizational policy affected patterns of usage in two contextually different cases.

To be clear, we do not measure the overall number of uses of offensive terms before and after the policy shift. Instead, we sample tweets that used specific terms before and after the change in policy to study whether message sentiment differs. We examine whether the use of such terms was influenced by Twitter’s attempts to reduce abusive content on the platform, in other words. Note that our design lacks a control group—the policy shifts we study applied to everyone on the Twitter platform—which precludes approaches such as difference-in-difference estimation. In this study we also focus on the short time period spanning one month before and after the policy shift. This limited time span helps ensure that any observed changes in Twitter content are less likely to be
attributable to intervening factors likely to occur during lengthier time periods. Also, we present analyses that focus on individual users who used derogatory terms to ask whether the policy shift affected a change in tweet content at the individual level (e.g., Flores 2017).

Using this data set, we do the following:

1. Compare the degree of negative sentiment in tweets targeting individuals on the basis of gender and/or race in a multivariate regression analysis before and after the enactment of Twitter policy changes.
2. Investigate and describe incidences of online harassment that target individuals on the basis of gender and/or race, and their network spread, before and after the enactment of Twitter policy changes.
3. Examine the pattern of retweets and likes of tweets that contain either sexist and/or a racist slur during the periods before and after Twitter policy changes, both of which can contribute to the network spread of a post.

**Methods**

**Data**

To test whether Twitter’s policy change influenced aggressive messaging, we selected a targeted approach in which we focus on one type each of a sexist and a racist slur. Data demands prohibited investigation of all types of sexist and racist tweets, so we investigated two of the most frequently used offensive words. We collected data directly from the Twitter application programming interface and scraped tweets that included either a gendered slur, *bi!ch*, or a racialized slur, *n!!ger*. We collected these tweets over a two-month period (November 17, 2017, to January 18, 2018), beginning one month before and concluding one month after the date of a publicized change in Twitter’s rules regarding abusive content (December 18, 2017). From this data set we extracted a random sample of close to 4 million tweets. After removing duplicated and empty tweets, the total sample size included 3,615,445 tweets. Although not a random sample of all tweets, we believe that our focus on tweets containing major gendered and racial slurs enables an especially stringent test of the effectiveness of Twitter’s rule changes. In effect, we test whether there was a detectable change in the sentiments captured even among tweets using terms that target women and racial minorities.2 As discussed in the following section, uses of these terms are sufficiently varied to make such a test meaningful.

Over the course of the two months (for a total of 64 days), we gathered an average of 56,491 tweets per day (1.6 percent of the total sample). The minimum number of tweets collected on a single day occurred on January 4, 2018, with 16,477 tweets (0.5 percent of the total sample), whereas the maximum took place on November 29, 2017, with 101,939 tweets (2.8 percent of the total sample). Tweets including the term *bi!ch* or *n!!ger* were not heavily concentrated on any single day but distributed fairly evenly over the two months.

**Sentiment Analysis**

We develop a classifier that we apply to the sentiment analysis of the content of the entire tweet message, which includes one of the two key words (Felmlee, Blanford et al. 2020; Felmlee, Inara Rodis et al. 2020; Zhang and Felmlee 2017). The sentiment tool is a variation of the VADER (Valence Aware Dictionary and Sentiment Reasoner) classifier. VADER is “a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.” It is fully open-sourced under the [MIT License](https://github.com/cjhutto/vaderSentiment). To improve the overall accuracy of the sentiment score, we compared the performance of VADER with that of several commonly used packages. In the end, we used a combination of scores from the top three classifiers, which included VADER and two others, afinn and bing, which are lexicon packages available in R. Moreover, we updated the lexicon to include derogatory terms relevant to incidences of harassment or aggression and translated the score into a scale ranging from −4 (most negative) to 4 (most positive). We selected this scale to replicate the original lexicon scoring of the VADER classifier and to standardize scoring across our final classifier. The key term *n!!ger* received an individual score that was more negative than that of the term *bi!ch*, according to our classifier. The sentiment score depended on the content of the entire tweet, however, not only the key word. We compared the sentiment scores gathered from four human coders on a sample of 400 tweets with the scores obtained by our final classifier. Our sentiment classifier performed well, with overall F1 scores of .688 (micro) and .690 (macro). These F1 scores represented an improvement over scores from VADER alone or from other combinations of common classifiers.

Unsurprisingly, most tweets containing our key terms are quite negative in sentiment. At the same time, there exists a

---

1. This feature of our analysis is possible only because we focused on tweets containing the slur terms rather than an overall sample of all tweets. Collecting such an overall sample would require using Twitter’s real-time “firehose,” and amassing tweets in real time would not allow us to track the network spread of those tweets over time. In contrast, our backward-looking, key term-based search allowed us to document which tweets received more likes and retweets.

2. Unfortunately, we cannot determine how many tweets we are missing simply because they were deleted either by users or by Twitter administrators. This limits our ability to identify precisely the mechanism by which the average sentiment of the tweets in our data changed after the policy change. We discuss this constraint further in the “Discussion” section.
range of sentiment in the data set, and not all messages are negative. Tweets that are positive, despite containing slurs, often reflect attempts by the user to “reclaim” negative terms by attaching a more positive meaning that would be recognizable to other in-group members (see Butler 1990). In exploring the online use of the term *bi!ch*, for example, Felmlee et al. (2018) found instances in which the gendered slur *bi!ch* is deliberately “resignified,” or “reclaimed,” on Twitter as an act of defiance and self-determination for women of color. Thus, tweets that score as positive in sentiment, even though they contain derogatory terms, often represent alternative or unconventional uses of slurs. Moreover, similar to what Pascoe and Diefendorf (2019) found in their study of tweets containing “no homo,” some tweets in this sample use racist and/or sexist slurs in a self-protective or boundary drawing fashion, where the use of a slur separates the user from other sexist or racist individuals, but the tweet itself is positive. In this project, we find some examples of such messages with “reclaimed” and “resignified” gendered or racial insults, as illustrated subsequently.

**Multivariate Approach**

To test the significance of Twitter’s policy against abusive messaging, we conduct ordinary least squares (OLS) regression analyses on the Twitter data and test the robustness of our findings using several other statistical approaches. The primary predictor, after policy enacted, is a binary indicator taking a value of 1 if the tweet is published after Twitter’s new policy implementation date (December 18, 2017) or 0 if it is published prior to this date. Tweets published on the date of the policy activation are considered to occur after the policy is enacted and scored as a 1. The dependent variable is the sentiment score of a given tweet (−4 to 4).

In addition to the main predictor, after policy enacted, further models also include several controls. To consider whether the length of a tweet makes a difference in the overall sentiment of the message, we include two controls. First, text character length is a continuous measure recording the number of characters in the tweet. Second, expected tweet length is a binary measure noting whether the total sum of characters in a tweet is between the minimum number of characters necessary to include the keyword (5 characters for tweets scraped with the term *bi!ch* and 6 characters for tweets scraped for the term *n!!ger*) and the total number allowed by Twitter. Tweets with messages containing between 5 or 6 and 280 characters were marked as 1.

To control for the influence of the time of day on tweet sentiment, we incorporate a variable, hours since midnight, for the hour of the day that the tweet was published. Golder and Macy (2011) found that tweet sentiment tends to vary over the course of the day. As previous research suggests a peak in tweet activity during midday (Cornell 2013), we also include a squared version of the time variable, (hours since midnight). Furthermore, because we collected the data over five major American holidays, we also add individual binary controls for whether tweets were published on a holiday (Thanksgiving, Christmas, New Year’s Eve, New Year’s Day, or Martin Luther King Jr. Day). In the models presented here, we include five variables, one for each holiday; tweets are marked 1 if the tweet is published on the respective holiday.

Finally, because of our interest in the possible network spread of deleterious tweets, we also include variables tracking the number of retweets, logged retweets, and likes, logged likes. We use a natural log transformation of the predictors because of the highly skewed distribution of both retweets and likes (i.e., the vast majority of tweets have no retweets or likes).

**Descriptive Results**

Our final sample is composed of just over 3.6 million tweets. The average tweet in the sample was published around 1:31 p.m., or just after midday (see Table 1). This matches previous research suggesting that tweets are most abundant around 1 p.m., after lunch or during afternoon work breaks (Cornell 2013). Within the sample, the average number of retweets was 2.46, and the average number of likes was 6.89. However, the modal numbers of retweets and likes were both zero. Thus, the vast majority of tweets were not further spread (retweeted) or approved (liked) by other users. Given that our sample consists of tweets containing racialized or gendered slurs, the average sentiment score of a tweet in this sample is negative, with a mean of −2.35. Notably, tweets on or around holidays (Thanksgiving, Christmas, New Year’s Eve, New Year’s Day, and Martin Luther King Jr. Day) had slightly less negative content on average. Tweets on those five days had an average score of −2.24.

**Illustrations of Tweets**

Next, we present illustrations of tweets that represent a range of sentiment scores in Table 2. For example, the first four...
Table 1. Description of Data Sample.

|                          | All Data |                          |                          |                          |
|--------------------------|----------|--------------------------|--------------------------|--------------------------|
|                          | Mean     | SD                       | Mean                    | SD                       |
| Date collected           | 12/17/2017 | 19.7                    | 11/29/2017              | 8.6                      |
| Time collected           | 13:31:36 | 643:49                  | 13:34:58                | 639:45                  |
| Number of retweets       | 2.46     | 232.45                  | 2.51                    | 228.79                  |
| Number of likes          | 6.89     | 596.29                  | 6.86                    | 602.13                  |
| Character length of tweet| 74.82    | 57.75                   | 73.94                   | 56.92                   |
| Sentiment score          | −2.35    | 1.35                    | −2.36                   | 1.34                    |
| Number of tweets (percentage of total) | 3,615,445 (100) | 1,826,669 (51) | 1,788,776 (49) |

Note: Data were collected from the Twitter application programming interface. We calculate date collected by using the number of days since the first day of data (November 16). The mean date of the total data sample was 31.24, which translates to December 17. The standard deviation for date collected refers to a number of days. We calculate time collected by using the number of hours since midnight. The mean of the time data is 13.527, which translates to 13:31:37 or 1:31 p.m. The standard deviation for time collected refers to a number of hours.

Table 2. Examples of Twitter Messages and Their Scores.

| Tweet Content                                                                 | Score  |
|-----------------------------------------------------------------------------|--------|
| The n!\$ger girl is ugly as sh!t                                          | −4.000 |
| well you are a terrible fu!!ing person so unfollow me you ugly a\$ stale a\$ moldy a\$ little b!ch | −4.000 |
| I hope somebody kills this dumb stupid b!ch tonight . . .                    | −4.000 |
| B!ch stop trolling for comments and likes. You know go\$m well you tweeting stupid sh!t just for the attention on twitter. Grow up URL | −4.000 |
| I tweet n!!gers Fu!!ing dead bodies and nothing is happen to my Twitter so URL . . . | −3.640 |
| A Bi\$ch Would Say ANYTHING To Make Me Look Bad [Woman Emoji with Hand across her Face] LAWWWDDDD | −3.514 |
| Bitch stop lying [Crying/Laughing Emoji 😂] URL | −3.166 |
| I wish this n!!ger would stop taking the football out of the endzone | −2.693 |
| Watch yourself b!ch                                                         | −2.324 |
| Took them too long to get that n!!ger outta there                          | −1.350 |
| Karma really is a b!ch. Better watch your back, I might be your karma        | −0.915 |
| [Crying/Laughing Emoji 😂] [Crying/Laughing Emoji 😂] you actually ‘censored’ b!ches lmao you’re a special kid [Sunflower Emoji] [Leaf Emoji] URL | 0.828 |
| HEY B!CHES I MADE IT TO FOURTEEN. IM SO PROUD OF MYSELF. ITS YOUR GIRLS BDAY | 0.312 |
| USER thanks my n!!ger God bless you                                         | 1.011  |
| Wow Karen really is the smartest b!ch out there                             | 1.551  |
| happy new year my n!!ger USER all the best 2018                             | 2.004  |
| @USER here comes the talented and amazing b!ch. Ty for being the hilarious and kind person ik. I wish u all the best and I’ll always be here if u need me [Purple Heart Emoji ❤] IMAGE | 3.633 |
| Being a mixed mulatto is fun and entirely thanks to the superior genes and phenotypes and uncanny ability to understand both 1950’s bebop and progressive death metal, but being able to use the word n!!ger whenever I want is the true perk. | 3.719 |

rows of the table contain four tweets that receive the most extreme negative sentiment score of −4. One of these expresses hope that someone will kill “this dumb stupid b!ch,” while another severely insults a woman’s looks: “The n!!ger girl is ugly as sh!t.” Other tweets are more neutral in sentiment, with scores close to 0, one of which contains contrasting emotions (e.g., laughing/crying emojis); the positive and negative emojis offset each other and result in an overall sentiment score that is relatively neutral. Finally, there are examples of tweets that are exceptions to the rule that slurs are always negative and in which the content of messages with the term b!ch or n!!ger appears to be positive. One example of such an upbeat tweet is in the third to last set of tweets in Table 2, in which the original message uses the term n!!ger while greeting another user “Happy New Year”; this tweet’s score is 2.00. In another instance, a person is described as an “amazing b!ch,” and the tweet receives a highly positive score of 3.63. Both tweets illustrate “resignification,” whereby people use and reclaim negative terms for use in more positive contexts.
Aggressive Tweets and Gender

To study online harassment and gender, we chose a popular, negative keyword, bi\'ch (e.g., Felmlee et al. 2018) to scrape from the Twitter database. As noted earlier, this term is gendered in that the use of the term bi\'ch often attacks feminine characteristics. Our final sample of tweets that contains a mention of this keyword was 3,540,669 tweets.

In our first analysis, we highlight social networks of Twitter “conversations” that emanate from an original tweet that uses the term bi\'ch in an insulting manner to target a particular woman. In one example (Figure 1, left visual), we examine the following tweet: “Bi\'ch stop lying 😂.” This brief tweet accuses an anonymous female target of lying and labels her a “bi\'ch.” Moreover, even without identifying the target of this attack, 10 others retweet or like the original aggressive post. In the visual figure, two network edges, or lines, represent users who both retweet and like the original post, and the other eight represent 8 users who like the original tweet.

In the second example (Figure 1, right visual), we illustrate a highly negative Twitter conversation, based on the tweet “well you are a terrible fucl!ng person so unfollow me you ugly as stale as moldy as little bi\'ch.” In this message, the originator of the message uses multiple insulting terms to attack two individuals, culminating in calling each of them a “terrible” person and using curse words and affronts to amplify the abuse. The perpetrator of this post also asks the target to “unfollow” them on Twitter. Both the nastiness of the language in the tweet and the request to be “unfollowed” are noteworthy. The latter is relevant, because unfollowing someone effectively cuts off the person not only from the bully but also from the network of other followers. Another significant aspect of this case is the relatively large network of conversation that emanates from the original post. In total there are 215 individual users who participate in this network. Interestingly, 114 of these users (53 percent) are marked private, meaning that a casual observer can see that they retweeted or liked messages in the conversation but cannot view the message content or further details about these users. Both gender examples occurred before the policy change on abusive content by Twitter.

Aggressive Tweets Targeting Race

Next, we analyze messages targeting race by collecting tweets that include the term n!!ger in the tweet (74,776 tweets or 2.1 percent of the total sample). In the first example (Figure 2, top
visual), we consider the following tweet: “AYE LIL N!!GER GET IN THIS FUC!!ING CALL RIGHT BEFORE I GO BUY A SNAKE, DRIVE DOWN TO TEXAS, FIND YOUR HOUSE AND PUT THAT B!CH NEAR YO A!S CRACK.” In this instance, the person sending the message uses an offensive, racial slur (n!!ger) as well as curse words to make the point. The social network surrounding this message consists of five people who continue to abuse one another in subsequent tweets. Furthermore, in this relatively small conversation network, it appears that the Twitter users know one another, and therefore they are using insulting language not to affront strangers but to emphasize a point aggressively with acquaintances. Not all offensive messages target strangers.

In the second example (Figure 2, bottom visual), we illustrate a message that contains the racist slur p!ki, a term used to attack individuals from Pakistan, Asia, or from other non-Western European backgrounds. The selected tweet we illustrate is “You fat ugly p!ki b!ch, lol you have facial hair. You literally have a mustache, hairy arms, and legs. Fuc!!ng die you b!ch — my hair keeps me warm it’s cold out here.” In the tweet, the original user is responding to an attack offsite and seems to be venting resentment. As can be seen in the figure, a total of 45 individuals like, retweet, or reply to the original user’s post, and all of them act in support of the message, despite its vitriolic content. Here, again, we see the extensive involvement of the social network in responses to online aggression. In these two instances of racial targeting, the former occurs before the policy change. However, the latter was published after the change, demonstrating that negative messages still can occur following the shift in organizational rules.

**Role of Policy in Aggressive Tweets: Multivariate Analyses**

Next, in a series of multivariate regression analyses, we formally address the question as to whether a shift in policy significantly affects the degree of sentiment negativity in our full sample of tweets. To do so, we compare sentiment scores before and after Twitter enacted new policies intended to stem abuse on their platform.
As can be seen in Table 3, rule changes on the part of Twitter are related significantly to message sentiment. Given the size of our sample, it is not surprising that we report coefficients that are highly statistically significant; more important, however, is that the coefficient for the policy shift is positive in sign. This is notable because it cuts against the expectation that reactance dynamics might motivate users to tweet even more negatively when they perceive Twitter as trying to control their behavior. For example, in the first of our three models (the “null model”), where the only predictor is after policy enacted, the main predictor variable is positive ($\beta = .001$, $p < .001$); this means that tweets after the date of the Twitter changes intended to reduce abuse are slightly more positive in sentiment than those published prior to that date.

Next, in the “partial” model (Table 3, column 2), we include controls for the length of the text, the expected tweet length, the time of the tweet (i.e., hours since midnight and hours since midnight squared), and whether the date represents a holiday. For example, in the first of our three models (the “null model”), where the only predictor is after policy enacted, the main predictor variable is positive ($\beta = .001, p < .001$); this means that tweets after the date of the Twitter changes intended to reduce abuse are slightly more positive in sentiment than those published prior to that date.

Retweets and Likes

An important element of messages on Twitter is that people can like and retweet others’ tweets, thus helping spread Twitter content through the social media network. In the third regression model, the “full model,” (Table 3, column 3), therefore, we also investigate the role of retweets and likes in message sentiment by adding two additional variables to our previous “partial” model. Findings for the previous control variables continue to be similar to those in the partial model, with only small shifts in coefficient size. However, in the full model we see that both retweets and likes significantly influence message sentiment, but in opposite directions. The variable “logged retweets” relates negatively to tweet sentiment ($\beta = -.068$, $p < .001$), whereas “logged likes” has a positive relationship ($\beta = .061, p < .001$). Tweets that spawn retweets, in other words, are likely to be more negative than those without retweets. Those that gather likes, on the other hand, tend to be more positive. The effect of the policy change remains statistically significant and positive in this expanded model ($\beta = .012, p < .001$).

Table 3. Three Models Predicting Tweet Score on the Basis of Characteristics of Tweet.

|                      | Null Model | Partial Model | Full Model |
|----------------------|------------|---------------|------------|
| Intercept            | −2.361***  | −2.177***     | −2.149***  |
| After policy enacted | .023***    | .012***       | .011***    |
| Scraped using bitch  | −.100***   | −.113***      | −.114***   |
| Text character length| −.000***   | −.000***      | −.000***   |
| Expected tweet length| −.117***   | −.124***      | −.124***   |
| Hours since midnight | .005***    | .005***       | .005***    |
| (Hours since midnight)$^2$ | −.000***   | −.000***      | −.000***   |
| Thanksgiving         | .135***    | .133***       | .133***    |
| Christmas            | .135***    | .132***       | .132***    |
| New Year’s Eve       | .188***    | .185***       | .185***    |
| New Year’s Day       | .138***    | .133***       | .133***    |
| Martin Luther King Jr. Day | −.002      | −.002         | −.002      |
| Logged retweets      |            | −.007***      | −.007***   |
| Logged likes         |            |               | .012***    |
| Number of observations| 3,615,445  | 3,615,445     | 3,615,445  |
| Multivariate $R^2$   | .00007     | .00144        | .00248     |
| Adjusted $R^2$       | .00007     | .00144        | .00248     |
| AIC                   | 12,453,336 | 12,448,404    | 12,444,658 |
| BIC                   | 12,453,375 | 12,448,574    | 12,444,854 |

Note: AIC = Akaike information criterion; BIC = Bayesian information criterion.

*p < .001.

5Clearly, these associations cannot be interpreted as a causal effect of the predictor variables on the dependent variable, as the sentiment of the tweet is already determined by its content before that tweet gets liked or retweeted.
Gender and Race

We also perform further analyses to test whether gendered or racialized tweets exhibit divergent patterns with regard to the findings presented above. We begin by splitting the sample into two: one containing tweets scraped using the term *bi!ch* and one with tweets scraped using the term *n!!ger*. We then run the same analyses described above to test if there were any differences by sample. Although the overall findings remain the same—tweets are more positive following the policy change—there are two slight differences. First, tweets published on Martin Luther King Jr. Day containing the term *n!!ger* are significantly more likely to be positive than on other days, although this holiday effect does not occur for the overall sample or for tweets with the term *bi!ch*. Second, the retweet pattern detailed for the whole sample reverses for the smaller sample of tweets including the term *n!!ger*. Retweeted messages containing *n!!ger* were significantly more likely to be positive, rather than negative, in sentiment. According to additional investigation, this trend appeared to be due to an idiosyncratic, well-publicized event involving the N-word during the data collection period.

Following Individuals’ Tweets over Time

In supplementary analyses (see Supplementary Table 2), we also examine the degree to which individual user behavior is influenced by the policy change. We analyze this first by conducting fixed-effects regression models to control for individual variation. We use a sample of users who tweeted at least once both before and after the alteration in rules, thus allowing us to test a difference in behavior over this period. In this sample, there were 312,869 unique individuals who used the B-word once before and after the change (20.15 percent of all individual users in the total data set of users who...
use the B-word) and a relatively small group of 4,142 unique users who used the N-word in this same pattern (9.49 percent of all tweets scraped containing the N-word). In the final models, individuals’ tweets after the policy remain significantly more positive than tweets before the policy (although in complex models with multiple interactions the positive effect is not always significant).

Additionally, we conduct supplementary OLS regression models (see Supplementary Table 3) to control for individual variation. In this case, we include a binary flag for whether a user tweeted both before and after the policy and a count variable for the number of tweets a user had over the whole period. In these models, the coefficient for the policy change continues to be significant and positive. The total number of tweets a user sent also is associated with slightly more positive sentiment.

**Network Sentiment Change over Time**

To further illustrate the influence of Twitter’s policy change on the overall sentiment of tweets we visualize a subsample of tweets before and after the change (see Figure 3). We take a sample of tweets (those including both the term n!!ger and another user’s account in their original tweet) to specifically highlight how the anti-abuse rules may influence the practice of directly messaging other users. Within the visualization, we lock users in place (i.e., dots in Figure 3) before and after the policy to note how users tweeted using the racialized slur during these two periods. To emphasize the change in overall sentiment we colored the tweets as red, gray, or blue lines depending on their sentiment scores. Negative tweets, which received scores between −4 and −1, are depicted in red. Neutral tweets were assigned scores between −1 and 1 and are depicted in gray. Positive tweets, which received scores between 1 and 4, are depicted in blue.

Note that Figure 3 contains two large hubs (a user with a large abundance of direct mentions from other users) on either side of the network. These hubs illustrate that many users directly mentioned many of the same Twitter accounts in their posts using n!!ger. In addition, we see a reduced concentration of negative tweets (red lines) in the center of Figure 3, which is apparent when comparing the right visual (tweets after the policy) with the left (tweets before the policy). The reduction in negative tweets subsequent to the change suggests that Twitter’s anti-abuse policy has some visible positive effects. Given that this finding represents only a small fraction of the overall data, it is notable that the implementation of new rules can influence the behavior of even a limited portion of tweets.

An additional important finding visible in Figure 3 is that despite an overall positive influence of the anti-abuse policy, potentially damaging messages still spread in clusters of groups. These results reinforce the regression analyses finding on the basis of the overall sample of tweets regarding retweets, in which tweets with more retweets are likely to be negative. In other words, tweets containing slurs that spread and emanate through retweet threads are more negative in nature than tweets that are not being retweeted or viral in nature.

**Other Robustness Tests**

In addition to the models presented here using OLS regression on the total sample, we perform further analyses to test the robustness of these results. On the basis of the previous findings, we first explore the role of tweets published on holidays. In supplementary analyses, we drop all tweets published on holidays, and the overall findings remain the same, with the positive policy effects slightly larger in size (see Supplementary Table 4). We also try to control for bots and find that even when accounting for the presence of demarcated bots in the data set (e.g., accounts containing versions of the term bot), the positive finding remains relatively unaltered (see Supplementary Table 3). Moreover, we compare our results from this modeling approach with several others, including ordered logistic regression and partial proportional odds logistic regression, using both the Akaike information criterion and mean squared error of fitted values. Both the significance and the direction of the effects in the findings detailed above remain the same when using these other modeling strategies, but the overall fit is not as good as that for the analyses presented herein (see Supplementary Tables 5 and 6).

As a final robustness check to ensure that the policy effect was caused by the policy intervention rather than a spurious variable, we performed a series of analyses in which we randomized the date of each tweet (see Supplementary Table 7). For each model, we estimated the new effect of the policy change on the shuffled data and repeated this test 1,000 times. We then compared the replicated models with our empirical model and found no instances in which the policy predicted the randomized data. From these analyses, we find that it is exceedingly unlikely that the policy effect we observed was a product of chance, as none of the permuted data sets produced a coefficient as large as the one in the original, unshuffled data.

**Discussion**

In sum, using a sample of tweets with a common gendered slur and/or a racist insult, we see that a change in Twitter policy positively influences the sentiment score of a sample of more than 3.6 million tweets. The modest positive effect remains even after we incorporate a range of control variables

---

4If there are a large number of remaining bot accounts we have not identified, we would not expect this to confound our main findings; there is no particular reason to think that bot tweeting became more or less prevalent after the policy change was enacted.
in our models to measure the influence of tweet length, time of the post, holidays, retweets, and likes. Our finding is robust to an analysis in which we investigate the effect of days close in time to the policy change; to a model that compares sentiment in tweets posted by the same set of individuals prior to, and subsequent to, the shift; and to a comparison of 1,000 models in which we randomly shuffled the date of each tweet.

Our case studies, furthermore, reveal the way in which retweets and likes of aggressive messages can lead an initial nasty post to spread beyond the original target to involve others in a virtual online social network. The potential for harm arising from a single tweet, in other words, is magnified because of the networking environment inherent in social media. Note, too, that messages that are retweeted tend to be more negative than those that do not get sent forward in a retweet. This result means that malicious tweets are more likely to spread beyond the initial post and in this manner may have particularly damaging effects.

No research project is without limitations. In this instance, a study on the influence of an organizational rule change on negative messages using Twitter data must be attentive to issues relating to the sampling of data, the temporal window before and after the rule change, selection of key terms, measurement of sentiment, handling data from bots, treatment of special days (holidays), and alternative analytical strategies. Throughout our work we discuss many of these challenges and report on specific robustness checks that might influence our results. Nevertheless, our study addresses only a limited portion of what represents a wider, societal problem. Additional research is necessary to broaden the scope to examine, for example, other manifestations of electronic harassment, alternative policies, and possible long-term effects.

Our main objective was to identify whether Twitter’s policy change had an effect; however, our data provide less leverage on the question of how this effect occurred. One possibility is that the rule change precipitated a shift in norms that led users to avoid negative or aggressive tweeting. However, it is also possible that the policy change led to more negative tweets’ simply being removed from the platform by fiat, thereby making the average sentiment more positive. Yet another possibility is that users found ways to effectively “code” their tweets targeting women and African Americans in ways that our sentiment classifier might not pick up. Future work with different data will be needed to more fully pursue these questions of causal mechanism.

In conclusion, this research applies a novel approach to studying the dynamics of a version of online aggression. Twitter’s rule changes presented a timely chance to investigate whether organizational “anti-abuse” policies can stem hateful speech on social media. The opportunity presented by this policy change enabled us to test several social scientific arguments regarding the influence of organizational efforts to sanction adverse behavior, and our results have implications for these theories. For example, given that negative sentiment decreases, albeit slightly, following the new rules, this suggests that organizational interventions may be able to play a role in curbing undesirable actions on the part of platform users. Perhaps of most note, we do not uncover significant increases in negativity that could reflect defiant reactions to corporate rules seen as restricting online rights. Yet we should emphasize that instances of offensive messages remain after rule changes, as shown in our case studies. Although not entirely immune to corporate policy, online aggression remains a sticky societal problem and one lacking simple solutions.

On a broad level, our findings call further attention to the serious social problem inherent in online harassment and aggression within social media. Our study represents one of the first large-scale investigations in which problematic Twitter messages are tracked that contain slurs that target individuals on the basis of gender or race. Our research confirms, in an independent investigation, that certain organizational anti-abuse practices can avoid inflaming users who use sexist and/or racist terminology to react with heightened hostility. Such policies, instead, may begin to limit the negativity of these all too common, distressing online interactions.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Support for this research was provided by the National Science Foundation under grant 1818497. We also acknowledge administrative assistance from the Population Research Institute at Penn State University, which is supported by an infrastructure grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (P2CHD041025).

ORCID iDs
Diane Felmlee https://orcid.org/0000-0003-4211-0489
Daniel DellaPosta https://orcid.org/0000-0002-5307-413X

Supplemental Material
Supplemental material for this article is available online.

References
Baldassarri, Delia, and Guy Grossman. 2011. “Centralized Sanctioning and Legitimate Authority Promote Cooperation in Humans.” Proceedings of the National Academy of Sciences 108(27):11023–27.

Bartlett, Christopher P., Caroline C. DeWitt, Brittany Maronna, and Kaleigh Johnson. 2018. “Social Media Use as a Tool to Facilitate or Reduce Cyberbullying Perpetration: A Review Focusing on Anonymous and Nonanonymous Social Media Platforms.” Violence and Gender 5(3):147–52.

Bearman, Peter S., and Hannah Brückner. 2001. “Promising the Future: Virginity Pledges and First Intercourse.” American Journal of Sociology 106(4):859–912.
Bellmore, Amy, Angela J. Calvin, Jun-Ming Xu, and Xiaojin Zhu. 2015. “The Five W’s of ‘Bullying’ on Twitter: Who, What, Why, Where, and When.” Computers in Human Behavior 44:305–14.

Brans, Heleen, and Bert Scholtens. 2020. “Under His Thumb: The Effect of President Donald Trump’s Twitter Messages on the U.S. Stock Market.” PLoS ONE 15(3):e0229931.

Brehm, Jack W. 1966. A Theory of Psychological Reactance. Oxford, UK: Academic Press.

Butler, Judith. 1990. Gender Trouble: Feminism and the Subversion of Identity. New York: Routledge.

Chatzakou, Despoina, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. “Mean Birds: Detecting Aggression and Bullying on Twitter.” arXiv. Retrieved July 27, 2020. https://arxiv.org/abs/1702.06877v3.

Chaudhry, Ifran. 2015. “#Hashtagging Hate: Using Twitter to Track Racism Online.” First Monday 20(2). Retrieved July 27, 2020. https://journals.uic.edu/ojs/index.php/fm/article/view/5450.

Cornell, Scott. 2013. “Peak Times for Twitter Activity.” Chron. Retrieved April 2, 2018. https://smallbusiness.chron.com/peak-times-twitter-activity-62864.html.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingrid Weber. 2017. “Automated Hate Speech Detection and the Problem of Offensive Language.” Pp. 512–15 in Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM). Retrieved July 27, 2020. https://arxiv.org/pdf/1703.04009.pdf.

DellaPosta, Daniel, Victor Nee, and Sonja Opper. 2017. “Endogenous Dynamics of Institutional Change.” Rationality and Society 29(1):5–48.

DellaPosta, Daniel, Yongren Shi, and Michael Macy. 2015. “Why Do Liberals Drink Lattes?” American Journal of Sociology 120(5):1473–1511.

Ey, Lesley-Anne, Carmel Taddeo, and Barbara Spears. 2015. “Cyberbullying and Primary-School Aged Children: The Psychological Literature and the Challenge for Sociology.” Societies 5(2):492–514.

Faris, Robert, and Diane Felmlee. 2011. “Status Struggles: Network Centrality and Gender Segregation in Same- and Cross-Gender Aggression.” American Sociological Review 76(1):48.

Faris, Robert, and Diane Felmlee. 2014. “Casualties of Social Combat: School Networks of Peer Victimization and Their Consequences.” American Sociological Review 79(2):228–57.

Fehr, Ernst, and Simon Gachter. 2002. “Altruistic Punishment in Humans.” Nature 415:137–40.

Felmlee, Diane H., Justine I. Blanford, Stephen A. Matthews, and Alan M. MacEachern. 2020. “The Geography of Sentiment towards the Women’s March of 2017.” PLoS ONE 15(6):e0233994.

Felmlee, Diane, and Robert Faris. 2016. “Toxic Ties Networks of Friendship, Dating, and Cyber Victimization.” Social Psychology Quarterly 79(3):243–62.

Felmlee, Diane, Paulina Inara Rodis, and Sara Chari Francisco. 2018. “‘What a Bitch!’ Cyber Aggression toward Women of Color.” Advances in Gender Research 26:105–23.

Felmlee, Diane, Paulina Inara Rodis, and Amy Zhang. 2020. “Sexist Slurs: Reinforcing Feminine Stereotypes Online.” Sex Roles 83:16–28.

Flores, René D. 2017. “Do Anti-immigrant Laws Shape Public Sentiment? A Study of Arizona’s SB 1070 Using Twitter Data.” American Journal of Sociology 123(2):333–84.

Golder, Scott A., and Michael W. Macy. 2011. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures.” Science 333(6051):1878–81.

Gusfield, Joseph R. 1966. Symbolic Crusade: Status Politics and the American Temperance Movement. Urbana: University of Illinois Press.

Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, and Juan Camilo Cardenas, et al. 2006. “Costly Punishment across Human Societies.” Science 312(5781):1767–70.

Juvenon, Jaana, and Elishева F. Gross. 2008. “Extending the School Grounds?—Bullying Experiences in Cyberspace.” Journal of School Health 78(9):496–505.

Lenhart, Amanda, Michele L. Ybarra, Kathryn Zickuhr, and Myesha Price-Feeney. 2016. “Online Harassment, Digital Abuse, and Cyberstalking in America.” New York: Data & Society Research Institute, Center for Innovative Public Research. Retrieved July 27, 2020. https://www.datasociety.net/pubs/oh/Online_Harassment_2016.pdf.

Lowry, Paul Benjamin, and Gregory D. Moody. 2015. “Proposing the Control-Reactance Compliance Model (CRCM) to Explain Opposing Motivations to Comply with Organisational Information Security Policies.” Information Systems Journal 25(5):433–63.

Marwick, Alice E., and danah boyd. 2011. “I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience.” New Media and Society 13(1):114–33.

Massanari, Adrienne. 2017. “#Gamergate and the Fappening: How Reddit’s Algorithm, Governance, and Culture Support Toxic Technocultures.” New Media & Society 19(3):329–46.

Nesterkin, Dmitriy A. 2013. “Organizational Change and Psychological Reactance.” Journal of Organizational Change Management 26(3):573–94.

North, Douglass C. 1990. Institutions, Institutional Change, and Economic Performance. New York: Cambridge University Press.

Pascoe, C. J., and Sarah Diefendorf. 2019. “No Homo: Gendered Dimensions of Homophobic Epithets Online.” Sex Roles 80(3):123–36.

Patchin, Justin W., and Sameer Hinduja. 2012. Cyberbullying Prevention and Response. New York: Routledge.

Pew Research Center. 2017. “Online Harassment 2017.” Retrieved July 27, 2020. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.

Ranco, Gabriele, Darko Aleksovski, Guido Caldarelli, Miha Grcar, and Igor Mozetic. 2015. “The Effects of Twitter Sentiment on Stock Price Returns.” PLoS ONE 10(9):e0138441.

Şahin, Mustafa. 2012. “The Relationship between the Cyberbullying/Cybervictimization and Loneliness among Adolescents.” Children and Youth Services Review 34(4):834.

Smaldino, Paul E., Marco A. Janssen, Vicken Hillis, and Jenna Bednar. 2017. “Adoption as a Social Marker: Innovation Diffusion with Outgroup Aversion.” Journal of Mathematical Sociology 41(1):26–45.

Stermer, Glenn, and Diane Felmlee. 2017. “The Social Networks of Cyberbullying on Twitter.” International Journal of Technoethics 8(2):1–15.
Tsugawa, Sho, and Hiroyuki Ohsaki. 2015. “Negative Messages Spread Rapidly and Widely on Social Media.” Pp. 151–60 in COSN ’15: Proceedings of the 2015 ACM on Conference on Online Social Networks. New York: Association for Computing Machinery.

Tyler, Tom, Matt Katsaros, Tracey Meares, and Sudhir Venkatesh. 2018. “Social Media Governance: Can Companies Motivate Voluntary Rule Following Behavior among Their Users?” Facebook Research. Retrieved July 27, 2020. https://research.fb.com/publications/social-media-governance-can-companies-motivate-voluntary-rule-following-behavior-among-their-users/.

Wang, Wenbo, Lu Chen, Krishnaprasad Thirunarayan, and Aamit P. Sheth. 2014. “Cursing in English on Twitter.” Pp. 415–24 in Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. Retrieved September 18, 2019. http://corescholar.libraries.wright.edu/knoesis/590/.

Xu, Jun-Ming, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. “Learning from Bullying Traces in Social Media.” Pp. 656–66 in Proceedings of the 2012 conference Of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Retrieved June 16, 2016. http://dl.acm.org/citation.cfm?id=2382139.

Ybarra, Michele L., Kimberly J. Mitchell, Janis Wolak, and David Finkelhor. 2006. “Examining Characteristics and Associated Distress Related to Internet Harassment: Findings from the Second Youth Internet Safety Survey.” Pediatrics 118(4):e1169–77.

Zhang, Amy, and Diane Felmlee. 2017. “You *&#*%! Identifying Bullying Tweets.” Presented at the 2017 Graduate Exhibition, University Park, PA.

Zych, Izabela, Rosario Ortega-Ruiz, and Rosario Del Rey. 2015. “Scientific Research on Bullying and Cyberbullying: Where Have We Been and Where Are We Going.” Aggression and Violent Behavior 24:188–98.

Author Biographies

Diane Felmlee is a professor of sociology and demography at Pennsylvania State University with research interests in social networks, social psychology, gender, and close relationships. Her research concerns the social processes that contribute to the interpersonal connections of peer aggression, online harassment, friendship, and intimate relationships. Her recent articles appear in American Sociological Review, PLOS One, Social Psychology Quarterly, Journal of Quantitative Criminology, and Sex Roles.

Daniel DellaPosta is an assistant professor of sociology and social data analytics at the Pennsylvania State University. His research applies network analysis across diverse research contexts, with a core focus on the mechanisms giving rise to intergroup cooperation and conflict, political and attitudinal alignments, and social and economic organization. His previous work has been published in the American Sociological Review, American Journal of Sociology, Social Forces, Social Networks, and several other outlets.

Paulina d. C. Inara Rodis is a doctoral candidate in sociology at Pennsylvania State University, where she earned her MA in sociology and demography. She is also a 2020-2021 Pre-Doctoral Fellow for Excellence through Diversity at the Annenberg School for Communication at the University of Pennsylvania. Her research interests focus on the interconnection between social identities and close relationships, social networks, and social media. Her dissertation examines minority women’s perceptions of and responses to aggression online. She has published work in Sociology of Education, Advances in Gender Research, and Sex Roles.

Stephen A. Matthews is liberal arts professor of sociology, anthropology, and demography (courtesy geography and School of International Affairs), and since 2013 the director of the Graduate Program in Demography at Penn State. He has written conceptual, methodological, and empirical papers on the connections between people and place focusing primarily on health disparities and spatial inequality. His research has appeared in the leading journals of several fields—sociology and criminology, demography, public health, epidemiology, and geography—on topics including residential segregation, neighborhood change, risk-taking behaviors, physical activity, food environments, alcohol outlets, crime, fertility, mortality, access to health services, cancer care, veteran health, women’s health, and religious landscapes. He is currently editor of Spatial Demography, associate editor at Mathematical Population Studies, and a past coeditor of Demography (2016-2019).