Cyberbullying on Twitter:

The Role of Gender, Race, and Policy in Negative Messaging

Paulina Inara Rodis, Diane Felmlee, Daniel DellaPosta, and Stephen Matthews

Pennsylvania State University

PAA Submission – September 19, 2018

Twitter can be an aggressive and hostile environment. People post approximately 15,000 bullying-related tweets daily (Xu, et al., 2012), suggesting that close to 100,000 harmful messages reverberate online weekly. News accounts repeatedly highlight the outcry surrounding damaging messages on Twitter. Based on a nationally representative survey (Pew Research Center, 2017), approximately 41% of Americans report having personally experienced some form of online harassment, with 18% describing particularly severe behaviors such as physical threats and sexual harassment. Blacks (25%) and Hispanics (10%) were more likely than Whites (3%) to report being targeted because of their race or ethnicity, and about twice as many women (11%) as men (5%) recount being harassed online due to their gender (Pew Research Center, 2017).

Aggression and violence toward women in the media remains extensive and problematic, as found in numerous studies. Reoccurring themes include that females are more likely than males to be sexualized (e.g., Dill and Thill, 2007) and treated as objects (e.g., Stankiewicz and Rosselli, 2008). Racism also persists within popular media. In video games, for example, Black characters are frequently linked to violence, a pattern that can reinforce stereotypes that Blacks are violent (e.g., Yang, et al., 2014). Furthermore, the depiction of minorities within advertising

and other forms of media reflects dominant cultural stereotypes about these groups (e.g., Taylor, Lee, and Stern, 1995).

To help stem some of the pervasive, negative content online, Twitter is embarking on a set of widely publicized policy changes to clarify existing standards and facilitate more reporting and removal of abusive content. In this project, we examine the effect of these rule changes on negative messages through sentiment analysis of tweets sent before and after the policy shifts. This moment provides a unique and timely opportunity to examine how the occurrence and spread of derogatory messages is influenced by Twitter's attempt to exert greater social control over its users.

We examine the degree to which these rule changes influence the occurrence and network spread of aggressive messages on Twitter, focusing on messages related to gender, race, and sexual orientation. Do the organization's policies alter the prevalence of hostile messages? We propose to examine this question in three ways.

- Compare the degree of negative sentiment in tweets targeting individuals on the basis of gender, race, and/or sexual orientation, and their network spread, before and after the enactment of alterations of Twitter rules
- Examine the social network spread of tweets targeting individuals on the basis of gender, race, and/or sexual orientation before and after the enactment of alterations of Twitter rules
- Investigate the geographic distribution of negative sentiment to see if the policy change stems harassment in certain regions of the country more than others

BACKGROUND

Cyber Aggression

Bullying poses a serious national and international problem (e.g., Faris and Felmlee, 2014), and one that now extends well beyond face-to-face encounters into the world of digital and Internet communication (e.g., Xu, et al., 2012). Cyber aggression, which refers to intentional, online messaging with the aim of insulting or harming someone, poses a serious social problem, and one that extends worldwide. Both victims and perpetrators of digital aggression are at risk of a host of negative behavioral and psychological outcomes (e.g., Nansel, et al., 2001). Being the victim of peer aggression has particularly deleterious consequences and is associated with anxiety, depression, and poor academic performance (e.g., Faris and Felmlee, 2014; Nansel, et al., 2001; Willard, 2007).

Role of Gender and Race in Cyber Aggression

Racist, sexist and homophobic tweets remain readily accessible to the general public at any time of day. Previous research found that it took an average of between 24 seconds to one and one-half minutes to locate the first of thousands of derogatory, aggressive tweets. This average was based on a sample of 28 searches of assorted defamatory slurs and insults (e.g., "Ni!ger"; "Wh!re;" "Fa!got"), averaged over various times of the day and days of the week (Sterner and Felmlee, 2017). Furthermore, such negative tweets often were retweeted or "liked" by followers, thereby creating networks of cyberbullying that spread far beyond the original perpetrator and target (Francisco, Rodis, and Felmlee, 2017; Lawson, Rodis, and Felmlee, 2017; Zhang and Felmlee, 2017).

DATA

To consider the effect of Twitter's policy change regarding abusive content, we begin by scraping tweets that included 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 suggests that "bi!ch" is a particularly common, negative slur used on Twitter (Felmlee, Inara Rodis, and Francisco, 2018). Further, this use is often negative and used to demean a problematic feminine target. We collect data directly from the Twitter API, then take a random sample and clean the resulting data; the remaining sample totals just over 3.8 million tweets.

METHODS

Sentiment Analysis

In previous work, we developed a classifier that will be used in the sentiment analysis of the tweets themselves (Zhang and Felmlee, 2017). The sentiment classifier is a variation of the VADER (Valence Aware Dictionary and sEntiment Reasoner) classifier, "a lexicon and rule-based sentiment analysis tool that is *specifically attuned to sentiments expressed in social media* (fully open-sourced under the [MIT License]" (https://github.com/cjhutto/vaderSentiment)). In our classifier, we update the lexicon to include derogatory and other targeted terms and translated the score into a -4 (most negative) to 4 (most positive) scale. After applying the classifier to our sample, the average score of a tweet in this sample is -1.813. Due to the key, negative term used to amass the data we expect the sample to be negative in sentiment on average. We present examples of tweets and their scores in Table 1; note that there are examples of positive tweets in the data. For example, Table 1 contains two highly negative tweets, which

are scored as -4, one of which expresses hope that someone will kill "this stupid bi!ch." Other tweets are more neutral in sentiment, with scores closer to "0," one of which contains contrasting emotions (e.g., laughing/crying emojis). Finally, there are examples of tweets that are exceptions to the rule, where the term "bi!ch" is used in a positive manner in the data. In one such instance, a person is described as an "amazing bi!ch," *and the tweet receives a positive sentiment score of 2.64.*

Statistical Analysis

To test the significance of Twitter's policy against abusive messaging, we conduct ordinary least square (OLS) regression analyses on the Twitter data. The primary predictor was a binary indicator taking a value of '1' if the tweet was published after Twitter's new policy implementation date (December 18, 2017) or '0' if it was published prior to this date. Following Twitter's policy implementation, tweets published on the date of the police activation are considered to be after the policy is enacted and scored as a '1.' The dependent variable is the sentiment score of a given tweet.

RESULTS

Role of Gender in Aggressive Tweets

The total sample analyzed in this project includes the keyword "bi!ch." This term is gendered in that the use of the term "bi!ch" often attacks feminine characteristics. 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 quit lying

[Crying/Laughing Emoji 😂]." The original poster sends this tweet without identifying a particular target and yet twelve other posters still retweet or like the original aggressive post. While short, the tweet itself describes an anonymous female target as a "bi!ch."

In the second example (Figure 1, Right Visual), we illustrate a highly negative twitter conversation, based on the tweet: "well you're a terrible fuc!!ng person so unfollow me u ugly a!s stale a!s moldy a!s little bi!ch." In this tweet, the original poster aggressively asks another to unfollow them on Twitter. Both the aggression in the language of the tweet and the request to be "unfollowed" are significant. The latter is significant because following others on Twitter helps individuals to spread their messages, so being unfollowed effectively separates one's network from another. Within the original post the aggressive language uses several different insulting terms to attack two specific individuals, culminating in calling each of them a terrible person and ugly, stale, moldy, little "bi!ch." Both of these examples occurred before the policy change on abusive content by Twitter.

Role of Race in Aggressive Tweets

Within the original sample there were also tweets using racist slurs. In the first example (Figure 2, Top Visual), we consider the following tweet: "AYE LIL NI!GER GET YOUR A!S IN THIS FUC!!NG CALL RIGHT NOW BEFORE I GO BUY A SNAKE, DRIVE DOWN TO TEXAS, FIND YOUR HOUSE AND PUT THAT BI!CH NEAR YO A!S CRACK." In this instance, the use of the term "bi!ch" (referring here to a snake) is less significant than some of the other insults in the tweet. Further, in this smaller conversation network it is clear that the Twitter users know one another and therefore are using insulting language not to affront strangers, but strike a stronger pose with acquaintances.

In the second example (Figure 2, Bottom Visual), we illustrate an aggressive response to a different, previous cyberattack that does not occur on Twitter. The selected tweet we illustrate is "You fat gross p!ki bi!ch, lol you have facial hair. You literally have a mustache and hairy arms and legs. fuc!!ng die you bi!ch — my hair keeps me warm it's cold out here URL." In the tweet, the original user is responding to an attack offsite and seems to be venting her resentment. Those 45 individuals who like, retweet, or reply to the original user's post all act in support of the original poster despite its vitriolic content. In these two instances, the former occurs before the policy change and the latter was published after the policy change.

Role of Policy in Aggressive Tweets

Results from three models predicting the score of the tweets can be found in Table 2. In the null model, where the only predictor explaining the sentiment of the tweets is *After Policy Enacted,* the main predictor variable is positive and significant ($\beta = 0.0215$, $p < 0.001$). In the partial and full models, which include additional controls, the effect of the policy continues to be significant and positive. This means that tweets after the date of the policy on abuse are more likely to be positive than those published before the change in policy.

One explanation for this pattern would be that tweets simply became consistently less negative over the two-month window we examined, but that the policy change itself was not the key driver of this change. To assess this possibility, we ran a further set of analyses in which we iteratively "changed" the policy enactment date; if the *actual* policy change on December 18 drove the decreasing negativity, we would expect the largest regression coefficients to cluster roughly around that date. Results from this iterative process are visualized in Figure 3. True to expectation, the regression coefficient for the policy enactment variable becomes gradually

larger as we move closer to the actual policy change date, and begins to decrease shortly after this date (the brief lag before the decrease begins suggests that the policy change took several days to "sink in" before having clear effects). Based on this pattern, we can have greater confidence in attributing the decrease in negative sentiment to the policy shift itself rather than other unrelated processes.

An important element of messages on Twitter is that people can like and retweet others' tweets, thus helping to spread Twitter content. In the full model (Table 2), 'Retweets' and 'Likes' both are significant predictors of the final score of tweets. Due to the highly skewed distribution of both retweets and likes (i.e., the vast majority of tweets have no retweets or likes) we used a natural log transformation of the predictors to help with the interpretation of their respective coefficients. The coefficient for retweets was negative ($\beta = -0.0611$, $p < 0.001$) suggesting that tweets with more retweets are more likely to be negative in sentiment score. The coefficient for likes was positive ($\beta = 0.0524$, $p < 0.001$), on the other hand, suggesting that tweets with more likes are more likely to be positive in sentiment score.

FURTHER DIRECTIONS

The results we present here use data scraped directly from the Twitter API using the keyword "bi!ch." To further explore the influence of gender, race, and policy on cyber aggression on Twitter, we plan to run similar tests on other data sources. First, we will explore geolocated Twitter data over the same time period. This will allow us to consider how the policy change may vary over different geographic regions. Second, we will use data collected with different slurs, such as those pertaining to one's gender, race, and/or sexual orientation identities. Third, we will attempt to uncover any new or old proxy derogatory terms that are used to try and

escape notice of abusive policies. For example, recent coverage of political code words online – the use of search engines as opposed to group names, or candies as opposed to ethnic slurs – suggests some possible ways of avoiding detection by Twitter rules.

CONCLUSIONS

The prevalence of negative and aggressive messages on social media, and especially those targeting women and racial or ethnic minorities, has been a topic of both public and scholarly interest in recent years. One vexing question has been whether those who manage the online platforms where such messaging takes place have the ability to shift behavioral norms through top-down policies aimed at preventing abusive messages (e.g. Ksiazek, 2015; Massanari, 2017). Adopting a quasi-experimental approach comparing tweets sent before and after a widely publicized policy shift enacted by Twitter in December 2017, we find evidence here that policy shifts *do* have a measurable impact on user behavior. Tweets after the policy were more positive than before the change. In particular, the policy decreased the negative sentiment of messages containing sexist or racist content. These findings suggest that Twitter and other platform managers can be more than passive observers of cyber-aggression, and that policy shifts aimed at protecting users from abusive messaging can have measurable impacts on user behavior.

SELECTED REFERENCES

Dill, Karen E. and Kathryn P. Thill. 2007. “Video Game Characters and the Socialization of Gender Roles: Young People’s Perceptions Mirror Sexist Media Depictions.” *Sex Roles* 57(11–12):851–64.

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.

Felmlee, Diane, Paulina Inara Rodis, and Sara Chari Francisco. 2018. “What a B!tch!: Cyber Aggression toward Women of Color” *Advances in Gender Research* 26:105-123.

Francisco, Sara, Paulina Rodis, and Diane Felmlee. 2017. *What a B⁉ Cyberbullying: Women of Color on Twitter*. Presented at The 2017 Graduate Exhibition, University Park, PA.

Ksiazek, Thomas B. 2015. “Civil Interactivity: How News Organizations’ Commenting Policies Explain Civility and Hostility in User Comments.” *Journal of Broadcasting and Electronic Media* 59(4):556-73.

Lawson, Jordan, Paulina Rodis, and Diane Felmlee. 2017. *Bigotry Takes to Twitter: Cyberbullying Towards African-Americans*. Presented at The 2017 Undergraduate Exhibition, University Park, PA.

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

McGoogan, Cara. 2016. “Internet Trolls Replace Racist Slurs with Codewords to Avoid Censorship.” *The Telegraph*.

Nansel, Tonja R., Mary Overpeck, Ramani S. Pilla, W. June Ruan, Bruce Simons-Morton, and Peter Scheidt. 2001. “Bullying Behaviors Among US Youth: Prevalence and Association With Psychosocial Adjustment.” *JAMA* 285(16):2094–2100.

Pew Research Center. 2017. *Online harassment 2017*. Retrieved from http://www.pewinternet.org/.

Stankiewicz, Julie M. and Francine Rosselli. 2008. “Women as Sex Objects and Victims in Print Advertisements.” *Sex Roles* 58(7–8):579–89.

Sterner, Glenn and Diane Felmlee. 2017. “The Social Networks of Cyberbullying on Twitter.” *International Journal of Technoethics (IJT)* 8(2):1–15.

Taylor, Charles R., Ju Yung Lee, and Barbara B. Stern. 1995. “Portrayals of African, Hispanic, and Asian Americans in Magazine Advertising.” *American Behavioral Scientist* 38(4):608–21. https://doi.org/10.1177/0002764295038004010

Xu, Jun-Ming, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. “Learning from Bullying Traces in Social Media.” Pp. 656–666 in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics. Retrieved June 16, 2016 (http://dl.acm.org/citation.cfm?id=2382139).

Yang, Grace S., Bryan Gibson, Adam K. Lueke, L. Rowell Huesmann, and Brad J. Bushman. 2014. “Effects of Avatar Race in Violent Video Games on Racial Attitudes and Aggression.” *Social Psychological and Personality Science* 5(6):698–704.

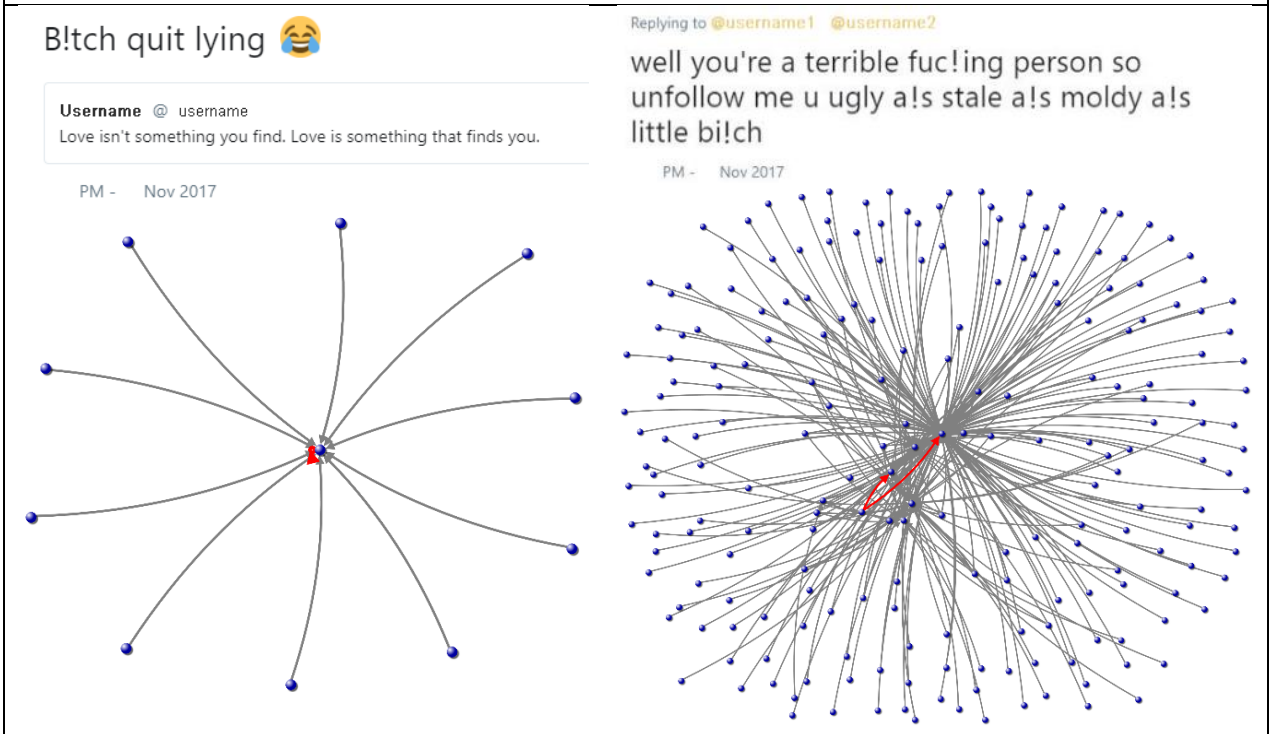
Zhang, Amy and Diane Felmlee. 2017. *You *&#*%!: Identifying bullying tweets*. Presented at The 2017 Graduate Exhibition, University Park, PA.

Tables and Figures

Table 1. Examples of Twitter Messages and their Scores	
Tweet Content	*Score*
well you're a terrible fuc!!ng person so unfollow me u ugly a!s stale a!s moldy a!s little bi!ch	-4.000
I hope somebody kills this stupid dumb bi!ch tonight ...	-4.000
Damn, you twitter bi!ches annoying	-3.573
Bi!ch stop trolling for comments and favorites. You know gotd!mn well you tweeting dumb sh!t just for the attention on twitter. Grow up URL	-3.569
A Bi!ch Will Say ANYTHING To Make Me Look Bad [Female Emoji with Hand across her Face] LAWWWDDD	-2.950
Bi!ch quit lying [Crying/Laughing Emoji 😂] URL	-2.339
Watch it bi!ch	-1.739
Karma is really a bi!ch. Better watch your back, I might be your karma	-0.685
[Crying/Laughing Emoji 😂] [Crying/Laughing Emoji 😂] you actually 'censored' bi!ches lmao you are a special child [Sunflower Emoji] [Leaf Emoji] URL	-0.069
HEY BI!CHES I MADE IT TO 14. IM SO PROUD OF MYSELF. ITS YA GIRLS BDAY	0.018
Wow Karen really the smartest bi!ch out here	0.936
@USERNAME1 here comes the talented and amazing bi!ch . Ty for being the funny and kind person ik. I wish u the best and I'll always be here if u need me [Purple Heart Emoji ❤] IMAGE	2.674
Note: Data scraped from Twitter API, then Scored Using Authors' Classifier Tweets are censored here, but not in original messages	

Table 2. Three Models Predicting Tweet Score Based on Characteristics of Tweet			
	Null Model	*Partial Model*	*Full Model*
Intercept	-1.82384***	-1.23381***	-1.23916***
After Policy Enacted	0.02154***	0.01682***	0.01581***
Text Character Length		-0.00214***	-0.00221***
Expected Tweet Length		-0.43474***	-0.44263***
Hours Since Midnight		0.01030***	0.01045***
Hours Since Midnight^2		-0.00055***	-0.00056***
Holiday		0.12365***	0.12044***
Logged Retweets			-0.06108***
Logged Likes			0.05241***
Mult. R^2	0.00008	0.01243	0.01329
Adj. R^2	0.00008	0.01243	0.01329
AIC	11576931	11531558	11528365
BIC	11576970	11531662	11528496
Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '•' 0.1 ' ' 1			

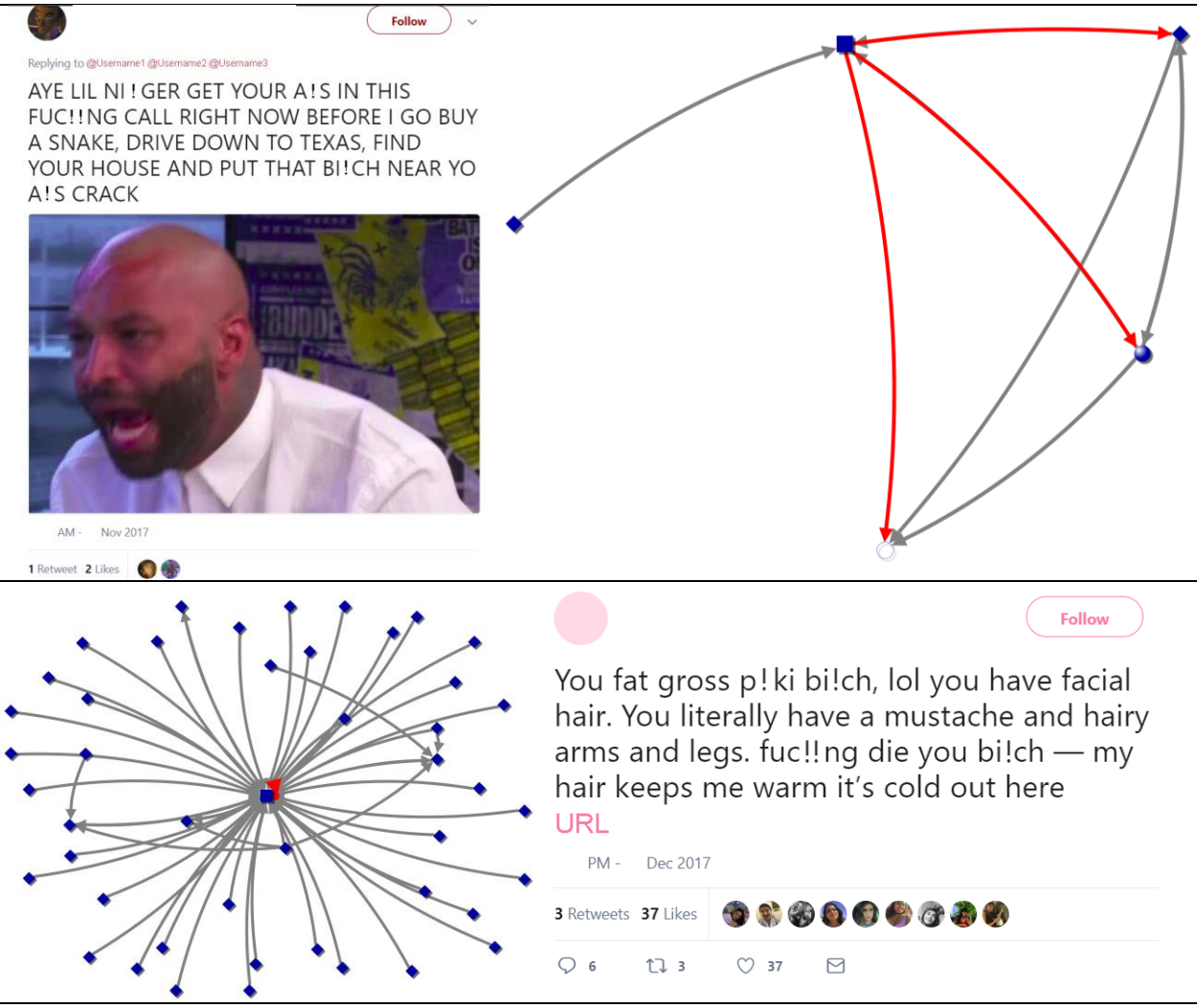
Figure 1. Network Visualizations of Twitter Conversations Based on Sexist Language



Note: Red Lines or Circles in each visualization illustrate the specific tweet highlighted within the conversation network

Tweets are censored here, but not in original messages

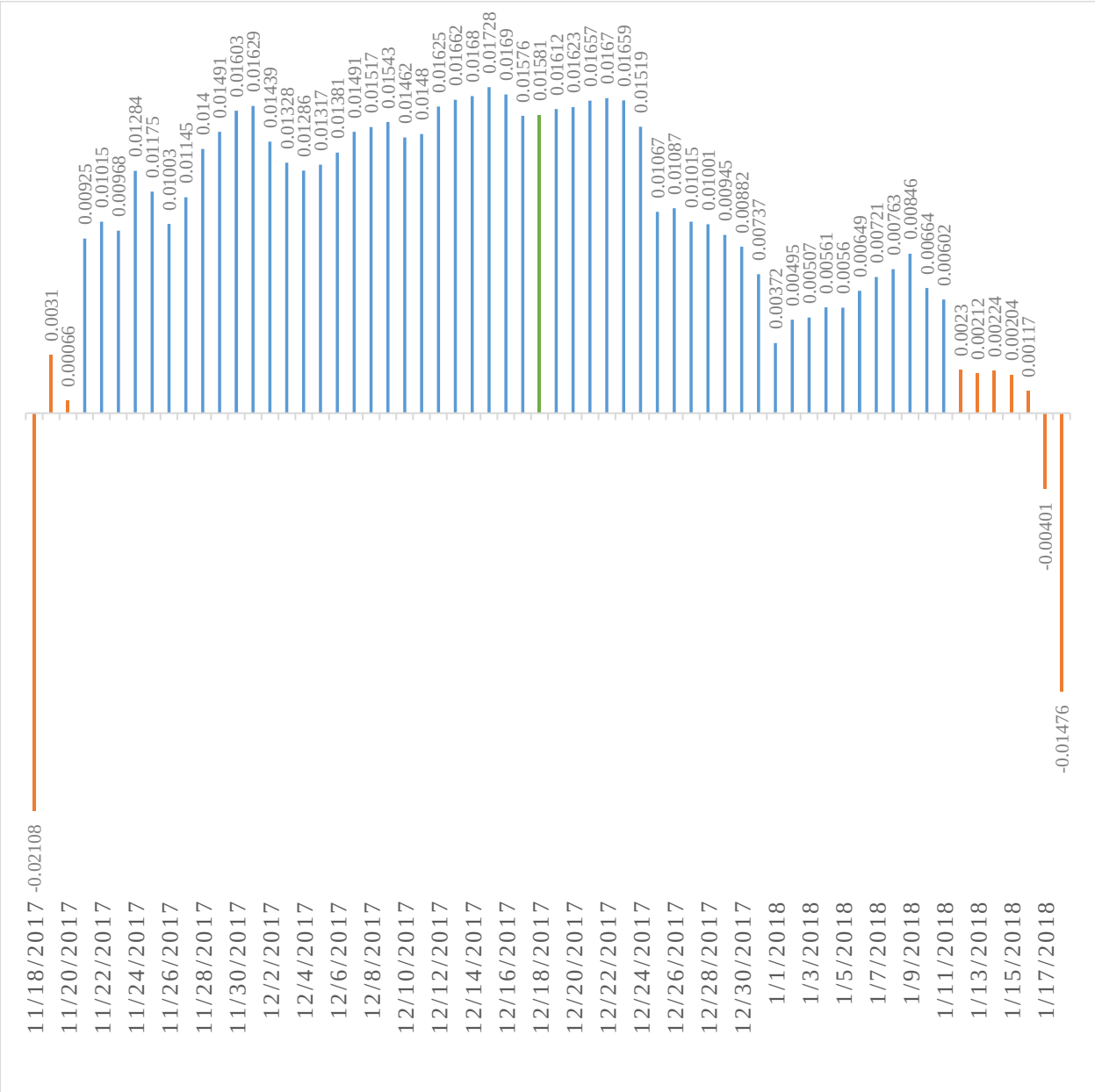
Figure 2. Network Visualizations of Twitter Conversations Based on Racist and Sexist Language



Note: Red Lines or Circles in each visualization illustrate the specific tweet highlighted within the conversation network

Tweets are censored here, but not in original messages

Figure 3. Prediction of Key Indicator Variable if Policy Enactment Date Changed



Note: Orange Lines represent coefficients that are not significant

Blue and Green Lines represent coefficients that are significant $p < 0.001$

The Green Line in the Middle represents the coefficient for December 18th