Rage Within the Machine: Activation of Racist Content in Social Media

Sebsatián Vallejo Vera[a,1]

[a]University of Houston

This manuscript was compiled on April 20, 2021

In social media networks, users engage and selectively propagate cognitively congruent frames. Racist content is among the frames users are exposed to. Unlike other frames, racist discourse is socially punished. In other words, racist discourse is a socially costly behavior among a menu of other less socially costly options. As such, people are unlikely to engage without a trigger. When will racist frames–a socially costly behavior–activate in social media networks? In this paper, I argue that social media users will engage with racist content when the status of the in-group is threatened. When the out-group threatens the status of the in-group, users will select frames that serve as markers to more starkly separate the in-group identity from the out-group. Racialized frames serve as these markers, and the threat to the in-group status makes racist content cognitively congruent. I provide evidence of this behavior by examining Twitter activity during the indígena protests in Ecuador in October 2019. I use a novel multi-step machine-learning process to detect racist tweets and show that pro-government users more actively engage when their status is threatened by pro-indígena users.

Racism | Social Media | Indígena Protest | Machine Learning | Ecuador

Social media has claimed a firm position in society, and today influences personal beliefs and political decisions across the world. Twitter and other social media networks have facilitated nearly instant connection and low-cost communication, including that of racist ideas and content. Yet while research has focused on the detection, characterization, and actors (Schmidt and Wiegand 2017; ElSherief et al. 2018) that engage in racist Twitter posts, we lack a systematic understanding of how this content draws social media users into a world of racist dialogue. This article explores the framing of racism in social media, paying particular attention to what and how individual users–rather than the media or institutions–choose to share, and what motivates users to promulgate these messages throughout their own networks. I ultimately find that threats to the status of the in-group play an important role in the propagation of racist social media posts.

Social media is a powerful tool to deliver and frame political narratives among voters (Barberá et al. 2015; Neumayer et al. 2016; Aruguete and Calvo 2018). In large part, these frames are developed through users' self-selection. Social media users will interact with and cluster around like-minded individuals (Himelboim et al. 2013), as the platform itself will subtly encourage these interactions thanks to the development of sophisticated algorithms. This results in social media users that frame events by collectively selecting or discarding content that is then impressed on the walls of like-minded peers. Within these social media bubbles, users "vote" on frames they find cognitively congruent, and discard frames they find cognitively dissonant (Aruguete and Calvo 2018).

These homogenous communities, resembling in-groups–users identify as members of social groups with shared identities (Keipi et al. 2017; Kakkinen et al. 2020)–are exposed to a marketplace of frames. Racist content is among the frames users are exposed to. Research has shown that, in political contexts, threats to the identity of the in-party (Amira et al. 2019) or the status of the in-party (Mason 2016) are drivers of out-group hate and anger. I extend this logic to the role of racism in social media and argue that in these homogenous communities of like-minded peers, racism is activated when the status of the in-group is threatened. Racist frames are particularly salient because they appeal to the identity of the in-group. When the out-group threatens the status of the in-group, users will select frames that serve as markers to more starkly separate the in-group identity from the out-group. Racialized frames serve as these markers, and the threat to the in-group status makes racist content cognitively congruent.

I provide evidence of this behavior by examining social media activity during the indígena (*indigenous*) protests in Ecuador of October 2019, a political crisis triggered by the decision of Lenín Moreno's administration to eliminate gas subsidies. As I show, racist content in social media networks is uncommon yet not rare. Despite the presence of racist messages during the span of the indígena protest, racist frames did not activate the pro-government community. As a socially punishable behavior–among other less costly frames–users were unwilling to engage with racist content. However, during events where the in-group status was explicitly threatened, such as Moreno acquiescing to the indígena movement's demands, the racial in-group identity of the pro-government community was activated, and racist frames became cognitively congruent.

Ecuador is an interesting test to explore racism in social media, a country with an organized and politically-active indígena community subject to historical manifestations of marginalization and racism. As a collective, the indígena community has challenged political power and gained political spaces (Van Cott 2008; Becker 2010), yet remains marginalized in the racially stratified Ecuadorian society (Hall and Patrinos 2004). The indígena political mobilization in an exclusionary state led to clearly-defined communities with conflicting interests

Significance Statement

In this research we analyze when racism discourse in social media is activated. We argue that users will engage with racists content when there is a threat to the group they belong to. We find support for our hypothesis in the behavior of Twitter users during the indígina protest in Ecuador in 2019.

The author is the sole contributor.

The author declares no conflict of interest.

[1]Correspondence should be addressed to Sebastián Vallejo Vera. E-mail: svallejoverauh.edu

and power relations (see Bretón and Pascual 2003), a reality that is manifested in our online social networks as well.

This study also presents a methodological contribution: an easy-to-implement strategy to detect racist tweets. Given the highly contextual nature of racist expressions, current dictionary-based, and machine learning techniques for detecting racism on the web perform poorly when applied to data in other languages or from different geographies. I use a combination of dictionary and semi-supervised machine learning (i.e. Google's Perspective algorithm) techniques to detect racism in the Ecuadorian network. This paper explains how to implement this approach in other contexts, and discuss the scope and limitations of this approach.

I begin by examining racism and social media framing, to then unpack the conditions that limit and heighten the proliferation of racist content in social media. I then introduce the particularities of the Ecuadorian case and present the Ecuadorian Twitter data and the multi-step process employed to detect racism. I use these data to show how users engage with racist content and the effect of events where the status of the in-group is threatened. I conclude with a discussion of the implications of this argument to the general study of racism within and beyond social media.

Social Media Framing and Racism

In social media networks, users tend to cluster around like-minded peers, what Himelboim et at. (2013) describe as selective exposure. Selective exposure occurs when individuals actively seek information that matches their beliefs, connecting with content that is cognitively congruent with their preferences and prior beliefs. Within these social media bubbles, individuals are exposed to information the is consistent with their beliefs, all the while deciding what content to propagate, and what content to not. In other words, users are selectively exposed to information that is cognitively congruent or dissonant with their preferences, and then decide whether to propagate this content across the network (Aruguete and Calvo 2018).

These social media bubbles are a marketplace of non-competing frames (Chong and Druckman 2007). Given the vertical configuration of social media networks, it is usually high-degree network authorities, users with significant number of followers, who are interested in framing social media events to their advantage. User are exposed to numerous frames, and "vote" among the choices. Cognitively congruent frames will propagate, cognitively dissonant frames will go unshared and unseen (Aruguete and Calvo 2018).

Racist content is among the frames users are exposed to. Like other frames in social media bubbles, racist frames will sometimes be cognitively dissonant to users and thus go unshared; other times it will be cognitively congruent to users and propagated. Unlike other frames, racist discourse is socially punished (Bonilla-Silva 2015). In other words, racist discourse is a socially costly behavior among a menu of other less socially costly options. As such, people are unlikely to engage without a trigger. When will racist frames–a socially costly behavior–activate in social media networks?

Homogenous communities formed in social media, particularly those in polarized environment, resemble in-groups with shared identities and social homophily (Keipi et al. 2017; Zollo et al. 2017; Kakkinen et al. 2020).[*] Social affiliations to gender, religious, and ethnic or racial groups promote in-group bias: greater attachments to and preference for members of the in-group (Tajfel 1981). In political parties, for example, these affiliations motivate members to advance the party's status (Huddy 2001).

However, in-group love is not reciprocal to out-group hate (Brewer 1999). Biased behavior towards out-group members (or out-party members) is not necessarily driven by desires to benefit the in-group (or the in-party). Denigrating the out-group does not advance the in-group or the in-party's status. Rather, in political contexts, threats to the identity of the in-party (Amira et al. 2019) or to the status of the in-party (Mason 2016) are drivers of out-group hate and anger. Furthermore, research has shown that anger, in particular, is a powerful political mobilizer (Groenendyk and Banks 2014), especially in strong partisans (Huddy et al. 2015).

I extend this logic to the role of racism in social media and argue that in these homogenous communities of like-minded peers, racism is activated when the status of the in-group is threatened. Racist frames are particularly salient because they appeal to the identity of the in-group. When the out-group threatens the status of the in-group, users will select frames that serve as markers to more starkly separate the in-group identity from the out-group. Racialized frames serve as these markers, and the threat to the in-group status makes racist content cognitively congruent. The mobilization of in-group users in social media is carried out by engaging with and propagating content (e.g. racist content), something that occurs with cognitively congruent frames.

From our proposed theory, the formulation of our hypotheses follow:

Hypothesis 1: Overall, users will find racist frames *cognitively dissonant* and will not engage with racist content.

Hypothesis 2: Events where the status or identity of the in-group is threatened will activate racist frames (i.e., *cognitively congruent*) among users.

The findings from Amira and colleagues (2019), Mason (2016), and Groenendyk and Banks (2014) are relevant in explaining activation of racist frames in social media, particularly when user frame political events. Social media has become a battleground where political narratives are delivered and framed among voters (Barberá et al. 2015; Bastos et al. 2015; Neumayer et al. 2016; Romero et al. 2011; Aruguete and Calvo 2018). The political network communities created by the selective exposure and dissemination of content and frames, often align with the political camps contending political power in "real life." Thus, the insights from the political psychology and political communication literature inform the engagement of users with racist content in social media. Mainly, that user will engage with racist content, one of many frames shared in social media bubbles, when the status of the in-group is threatened and racialized frames, especially those attacking the out-group, become cognitively congruent.

Racist Discourse in Social Media

Different from the race literature studying racial relations in the United States (Omi and Winant 2014) and Europe

[*] In addition to identifying as individual, people also identify as members of social groups to which they belong–i.e. in-groups–. People who identify as members of the "other" social groups are the out-group.

(van Dijk 1993), the literature in Latin America in general (Martínez-Echázabal 1998), and in Ecuador in particular (Roitman 2009), notes that race is a complex construction, due to mestizaje and the strong correlation between ethnic background, perceptions and auto-denomination of race, and class. In other words, how issues of race are framed, and how race itself is socially constructed, have their own geographical caveats. In Ecuador, as in other parts of the world, discursive racism is usually framed as an acceptance or tolerance of the out-group (e.g. indígena) by creating strict demarcations between the self and the "other." It is not surprising that racist language is both normalized and covert (Roitman and Oviedo 2017), even though these characteristics have different degrees and forms (De la Torre 1996). This is especially true since racism and race are often dismissed by individuals as explanations to racist behavior, and states structures and government representatives often pay mere lip-service to integration and ethnic identity. Yet, everyday patterns of behavior and speech, as well as the organization of the state, are configured in way that "indios" and "indígenas" are the subjects and objects of structural discrimination.

We would expect for online social spaces, such as Twitter, to replicate public and private racial discourses. Especially since online spaces amplify racist discourse, and unmask where racist discourse is produced and how it is reproduced (Eschmann 2019). The internet is a hybrid social space, at once public and private (Daniels 2012), where established and new forms of racism are facilitated (Back, 2002; Daniels, 2012; Nakamura, 2008). The user-generated communities in online platforms encourage intimate discursive interaction predicated in racial identity (Brock 2009; Daniels 2013). While overt racist discourse can either be socially reprimanded or legally punished, anonymity can lower the cost of racist behavior (Fox et al. 2015). Furthermore, research has shown that not only technical anonymity,[†] but also social anonymity–the user's *perception* of anonymity–can explain aggressive and hostile online behavior from a user's perceived freedom from social standards and sanctions (Christopherson 2007, Lapidot-Lefler and Barak 2012). Even in non-anonymous and moderated platforms (e.g. Facebook), users are comfortable expressing racist views (Chaudhry and Gruzd 2019).

There are diverse discursive manifestations of racism, both overt and covert.[‡] From a practical standpoint, I focus on overt forms of racist content that are easier to systematically detect than other forms of racism; overt racist forms that explicitly target someone because of their indigenous identity using negative and hurtful comments. But most importantly, in a country such as Ecuador, many platforms and institutions, classes and contexts, reproduce a "blanco-mestizo" racist ideology in often subtle and normalized patterns that many times the user is not aware or would not consider racist (Roitman and Oviedo 2017). With overt racism, the ambiguity is dispelled.

In the previous section I argued that user will engage with racist content when the status of the in-group is threatened. In what follows, I describe the Ecuadorian case, the data analyzed, and the methodological strategy used to test this hypothesis. First, I describe the indígena protests and the events surrounding the #ParoEcuador network in Ecuador. I then describe the Twitter network formed around the protests, and the well-defined pro-government and pro-indígena communities that emerged. I continue by providing an overview of the determinants of the production and reproduction of racists tweets. Finally, I provide evidence supporting the hypothesis proposed above by analyzing two political events where the status of the in-group, in the Ecuadorian case the pro-government community, was threatened.

The #ParoEcuador in Ecuador

On October 1 2019, Lenín Moreno, president of Ecuador, announced the elimination of gas subsidies. Two days later, the Unión de Transportistas (*Transportation Union*) announced a strike.[§] Two days later, the Confederation of Indigenous Nations of Ecuador (CONAIE), the largest indígena organization in Ecuador, followed suit, announced a strike, and started mobilizing their base towards the capital, Quito. By the time the indígena movement reached the capital, the president had moved the seat of government to Guayaquil and declared a state of emergency. In a polarized environment, pro- and anti-government media and protestors displayed contrasting sentiments towards the actions of both the executive and the protestors, as well as widely different accounts of violent incidents.

Posts related to the protests circulated extensively on Facebook and Twitter, the social media outlets with the largest user bases in Ecuador (Latinobarometro 2018).[¶] As state violence increased, many of the reports were initially broadcasted by online media sources, before being picked up by the more traditional outlets. For more than ten days the country was paralyzed, prompting mixed reactions from different groups. Labor unions joined the protests and various organizations, including universities, supported the indígena movement, especially as government violence increased. However, the protests also paralyzed an already faltering economy. Business representatives denounced the indígena protest, and praised the government for their position.

Beyond the role of social media in the Ecuadorian protest, there are important structural characteristics worth discussing. Racism in Ecuador is a "system of ethnic-racial dominance" historically rooted in European colonialism (van Dijk 2005), directed, in great part, towards the indígena population (Beck et al 2011). The indígena population has been marginalized by a state that has done little to support these communities, or grant them equal access to political spaces. Regardless, the indígena population has also managed to organize around a common banner (i.e. the "indígena" banner, despite the many, different, and sometimes conflicting nationalities) to demand and conquer important political and social victories (see Becker 2010). However, these victories have done little to change the racist ideology that permeates all levels of the state and society. Overall, despite their political activism and mobilization, indígenas and the indígena community have been marginalized and remain the main target of the national racist ideology.

[†] Technical anonymity refers to online interactions where there is no personal information of the user.

[‡] Many research frameworks examine more covert presentations of racism, including laissez-faire racism (Bobo et al. 1997), color-blind racism or no-difference racism (Bonilla-Silva 2003; Beck et al. 2011), and ventriloquism (Guerrero 1997). However, the discursive manifestation of these forms of racism are difficult to systematically identify.

[§] A couple of days after the announcement, the government negotiated a deal with the Unión de Transportistas and ended the strike.

[¶] Similar to other instances of social mobilization (Aruguete and Calvo 2018; Bastos et al. 2015, Gerbaudo 2012), media accounts of the 2019 protests in Ecuador (PBS 2019; Knight Center 2019) suggest that Twitter was part of the political battleground.

The #ParoEcuador Data

Between October 1 and October 24 2019, I collected three waves of Twitter data using the strings "paro" and "ecuador", two terms that are politically and racially neutral and were used by the government and indígena supporters alike. To collect this data, I connected *rtweet* (Kearney 2018) to Twitter's backward search application programming interface (API), gathering tweets in duration of the unrest in Ecuador. The data includes 2,425,239 posts by 85,249 unique Twitter users for the Ecuadorian case. Of this sample, 93% were retweets of an original tweet. I selected for our analysis only those accounts that participated multiple times and that were in the primary connected network.[‖]

The #ParoEcuador Network. As previously explained, selective exposure in social media leads to homogenous communities. During the Ecuadorian protests, we expect to find two well-defined communities: a pro-government community (in-group) and a pro-indígena community (out-group). In Twitter, this roughly translates into pro-government users mostly interacting (retweeting) with other pro-government users, and pro-indígena users interacting mostly with other pro-indígena users. In our Twitter network, each user is a node and an edge is created when a user H retweets user A.

To create the layout and identify the communities in the Ecuadorian Twitter network, I implement the following procedure: first, I load all the edges of the primary connected network with the author of the original tweet set as the authority (A) and the author of the retweet set as the hub (H), such that $H_{retw} \rightarrow A_{tw}$; second, I estimated a layout of node coordinates using the Fruchterman-Reingold (FR) forced-directed algorithm in R 3.5 *igraph* (Csardi and Nepusz 2006) and identified communities in the Ecuadorian network via random walk community detection. The FR algorithm facilitates the visual inspection of the network, communicating information about the proximity between nodes (data reduction pull) while preventing nodes from overlapping (force-directed push).

The random-walk community detection algorithm identified, as predicted, two primary communities in Ecuador: the pro-government network, which includes 36,579 nodes; and the indígena community network of 29,624 nodes. Figure 1 presents a basic FR layout of the Ecuadorian network. I describe the pro-government community with blue squares, and the indígena community with red triangles. The size of the nodes is proportional to the nodes' in-degree (authority), with larger nodes indicating users retweeted by a larger number of followers.

In Twitter, communities formed around political events and cleavages often have at their center political leaders or user strongly aligned to the leadership. Influential authorities in the 2016 United States election communities included presidential candidates @HillaryClinton and @realDonaldTrump; for the 2018 #Tarifazo networks in Argentina it was opposition leader @CFKArgentina (Cristina Fernández de Kirchner) and then president @mauriciomacr. In the Ecuadorian network during the 2019 indígena protests the pro-government community had at its center President @lenin, vice-President @ottosonnenh, and interior minister @mariapaularomo, as well as other prominent pro-government users. In the center of the pro-indígena community, there was the institutional account of the @CONAIE_Ecuador, and its president, @jaimevargasnae. Beyond public official or politicians, other influential users include media personalities, media outlets, and social media commentators.

The difference and sparsity of exchanges between each community give account of the polarization of the Ecuadorian network. Of all edges in the indígena community, 78.0% are with member of the same community (i.e. indígena community → indígena community), and only 6.5% with member of the pro-government community;[**] of all edge in the pro-government community, 91.6% are with member of the same community (i.e. pro-government community → pro-government community), and only 4.4% with member of the indígena community.

Detecting Racist Tweets. There is a large body of work dedicated to detecting hate speech in social media (Schmidt and Weigand 2017; ElSherif et al. 2018; Chatzakou et al. 2017), however, accurate and systematic hate-speech detection is a challenging task (Davidson et al. 2017). Despite the many advances on the topic, there are still limitations to the automatic detection of racist discourse, not the least of these, the contextual nature of racism (van Dijk 2005). While racist discourse will work to maintain existing power structures and racist ideologies, the language will evolve and shift, depending on the particularities of society and time. Thus, even if we were to take pre-trained models to detects hate speech, these would be useless in contexts different than the ones they were originally trained for.

To solve this problem, I adopt a multi-step classification approach, similar to that employed by ElSherif et al. (2018). I start by defining a racist attack towards a member of the indígena community or towards the indígena community in general as a "negative or hateful comment targeting someone because of their indigenous identity," a variation of the definition used by the Google's API *Perspective*.[††]. I use Google's API *Perspective*, a content moderating tool that is the industries' standard for automatic detection of toxic content in written comments. *Perspective* uses a convolutional neural network that scores the likelihood a text contains an identity attack.[‡‡] *Perspective* provides an identity attack score from 0 to 1, interpretable as the probability that a text will be perceived as an identity attack. I use a threshold score of 0.85 to create a dummy for whether a comment is an identity attack or not.[§§] The *Perspective* algorithm was trained to detect identity attacks on frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Since we are specifically interested in racist discourse directed towards the indígena community, I create a dictionary with key-phrases that identify the indígena community or members of the indígena community (i.e. *indígena, indio*). I keep only identity attacks (tweets) that contain "indígena",

[‖] I selected tweets with in-degree ≥ 2 and eliminated unconnected nodes.

[**] The algorithm identified a third, smaller community: the "opposition" community. The opposition community was made up of former president Rafael Correa and his political peers and followers, who also opposed the Moreno government. However, the opposition community formed a distinct cluster in the data from the indígena community, which is reflective of reality, as they had a common "enemy", yet did not coordinate with the indígena community. Note that most of the external dialogue of indígena community users was with opposition community users.

[††] Google's API *Perspective* considers an identity attack any "negative or hateful comment targeting someone because of their identity."

[‡‡] The model was built using millions of comments from the internet, using human-coders to rate the comments on a scale from "very toxic" to "very healthy", and using this large corpus as training data for the machine learning algorithm. See Wulczyn et al. (2017) for a comprehensive discussion on *Perspective*.

[§§] Lowering or raising the scores does not change the main outcomes, but it does change the accuracy of our model (see Appendix A).

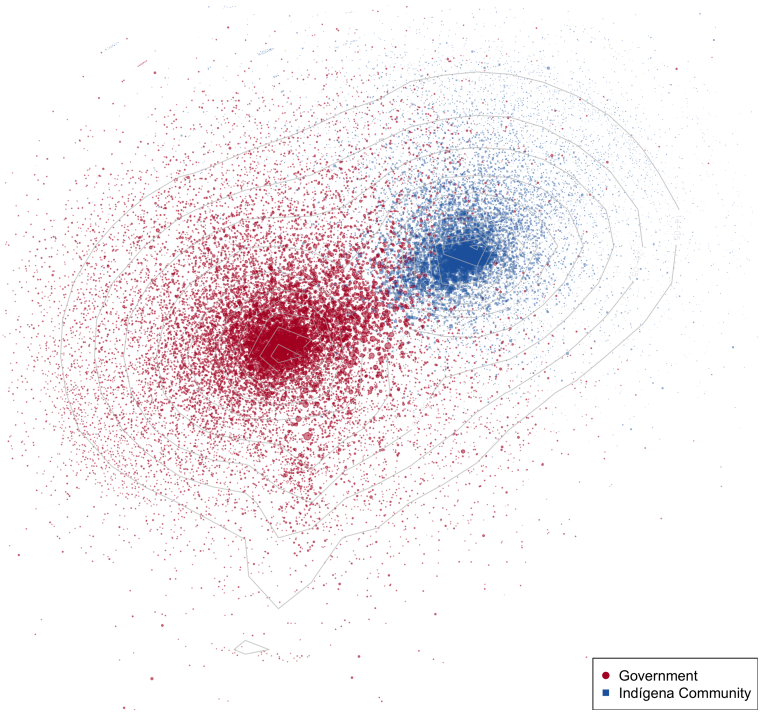


Fig. 1. Primary connected network during the Ecuadorian protests, between October 1 and October 24 2019. Red circles describe pro-government users. Blue squares describe users aligned with the indígena community.

the term most commonly used to refer to a member of the indígena community (e.g. "el/la indígena") or the indígena community in general (e.g. "los indígenas"). An alternative, more charged term to refer to indígenas is "indios." "Indio" (*Indian*) is often used by the blanco-mestizo population as a derogative identifier, and despite the long history of the indígena community reclaiming the term, it is still widely employed. However, the use of "indio" does not automatically reflect racism. Thus, I follow a similar procedure as before (i.e. detect identity attacks and keep those that include the term "indio"), but lower the threshold score to 0.75, given the charged nature of the term. Finally, I create a second dictionary with local forms of racist discourse that the algorithm is unable to detect. For example, during the protest in Ecuador, the phrase "*emplumados*," "feathered," was used to mockingly describe indígena leaders.[¶¶]

To check the internal validity of our measure, I hand-annotate a random subset of 1,500 tweets and compare the results to the ones obtained in our procedure.[***] In a sample limited to tweets from the pro-government community, the multi-step model obtains an F_1 score of 0.89, with a recall score of 0.85 and a precision score of 0.94, suggesting that our model is accurate and particularly good at avoiding false positives (Type-I error). In the hand-annotated sample there were no *explicit* forms of racism found in tweets produced by users from the indígena or opposition communities, yet the multi-step process detected racist tweets in those communities, a limitation discussed below. After applying these filters, I identify 1,371 (2%) racist tweets in the pro-government community. This multi-step process addresses both aspects of the definition for racist discourse: 1) the *Perspective* algorithm detects "negative or hateful comment targeting someone because of their identity", and 2) the key-phrases dictionary identifies the indígena community or an indígena as the target of the toxic tweet.

This multi-step process has some limitations. First, and most noteworthy, its reliance on the *Perspective* algorithm to identify racist tweets. Hosseini et al. (2017) showed drawbacks to the *Perspective* toxic detection system, mainly, under-detecting toxicity when key words are misspelled (e.g. words that signal toxicity) and sending false alarms for benign phrases denouncing toxic behavior.[†††] The latter was particularly problematic for tweets produced by users from the indígena or opposition communities. While the rate of racist tweets detected by our model in these communities was low relative to those identified in the pro-government community,[‡‡‡] all of these were incorrectly identified as being racist. These tweets were usually condemning and denouncing racist behavior and

[¶¶] For a complete list of terms, see Appendix A.

[***] For a detail recount of the process and comparative performance of our models, see Appendix A.

[†††] For example, "He said that they are idiots. That is not true." is considered highly toxic by the *Perspective* algorithm. In Appendix A we explain more in detail how the *Perspective* algorithm works and the reason behind false positives.

[‡‡‡] In the indígena community the algorithm detected 0.005% of racists tweets, compared to 2% from the pro-government community.

discourse from other users, political figures, media, police, etc., or pointing out racism in another user's tweet. Thus, I focus solely on users from the pro-government community.

Second, the *Perspective* algorithm is not able to capture more subtle forms of racist discourse, such as those that infantilize indígenas (Guerrero 1997) or use irony or wordplay to mock indígenas (Sue and Golash Boza 2013). These forms of racism can be found in all the communities, including the indígena and opposition communities. However, because of their subtlety, I am not able to systematically detect them through our model, so I limit the analysis to overt forms of racism. Exploring more subtle forms of racism is a promising avenue for further work. Finally, Twitter has various mechanisms that detect and potentially eliminate tweets with racist content.[§§§] Unfortunately, it is impossible to know whether a tweet was deleted or if it was deleted before the collection of the corpus (see Timoneda 2018).[¶¶¶] Overall, most of the limitations lead to underreporting of racist discourse in our corpus. However, underreporting leads to Type-II error, and I am comfortable with this possibility.

User Activation and Variables of Interest. The main variable of interest is whether a tweet is racist or not. We want to know who produces these tweets and, more importantly, how are these messages spread in our network. In social media, "acceptance equals propagation" (Aruguete and Calvo 2018). To operationalize "acceptance," and thus the spread of frames in a network, I follow Aruguete and Calvo (2018) and test whether latency increases or decreases for racist content, and how latency for racist content varies under different conditions. Aruguete and Calvo (2018) show that *time-to-retweet*, the number of seconds elapsed from the time a user (authority) posts a tweet to the time a second user (hub) retweets the same post, is a proxy for cognitive congruence. Users consuming messages in Twitter will propagate content from their community peers more quickly when these are cognitive congruent.[17]

In addition to the text of the tweet and the time of the tweet and the time of the retweet, I collected information on the user's screen name, followers, friends, and the status of the users' accounts (verified or not verified). From the network, I compute in-degree–the number of times a user has been retweeted–to identify high-authority users (i.e. users with significant number of followers) and low-authority users. Finally, I control for the effect of bots by including the ratio of followers to friends. In line with prior research, the Twitter data shows high degrees of concentration. In the #ParoEcuador network, less than 5.6% of the total accounts are responsible for 45% of the content that circulated in the network.

Sharing Racist Tweets in the #ParoEcuador Network

Before analyzing the (racist) reaction of pro-government users to threats to the status of the in-group, I present a more general snapshot of users interacting with racist content. Figure 2 describes user activity of the pro-government #ParoEcuador Network across time, labelled by key date. The red line represents overall tweets and the black line racist tweets. The activity of the network increased right after the Unión de Transportistas announced their strike, however, it was only after the CONAIE announced their strike that the racist content begun and continued for some time after the end of the strike, eventually waning as did the overall activity. The timeline shows how users became more active in the network as the strike intensified, and how racist activity did not necessarily follow the same pattern. In general, racist content was a latent frame in the pro-government community.

To more systematically explore the how users propagate racist content, I analyze the determinants for *time-to-retweet*. Higher values mean longer times between the original post of the tweet by the authority and the retweet of the post by the hub. I estimate a proportional Hazard Cox model, with unstandardized coefficients describing changes in the hazard rate of time-to-retweet. Positive coefficients indicate an increase in the hazard rate (faster *time-to-retweet*) and cognitive congruence with the frames; while negative coefficients indicate slower times and cognitive dissonance.

One important aspect to point out is that verified users[18] neither posted nor retweeted racists messages. This is likely a direct result of the cost public figures incur when engaging with racist content, something that less public users can get away with. Also note that many of the verified users are news outlets and government officials who often have, not only staff managing their social media accounts, but also more to lose if they were to publicly engage in overt forms of racism.

Table 1 shows the results from the proportional Hazard Cox model.[19] For an intuitive interpretation of the magnitude of the effect, consider the effect of our covariate of interest, "racist attack," among users of the pro-government community, which is negative and takes the value of -0.149 ($p \leq 0.01$) in Model 1. The exponentiated value of the coefficient (0.86) is the incidence rate, and can be interpreted as the (instantaneous) change in *time-to-retweet* when a tweet has racist content. One minus the incidence rate is 0.14, showing that the *time-to-retweet* for racist posts is about 14% slower. Thus, users are not eager to retweet racist tweets posted by their own authorities, suggesting that either the content of racist messages produce a cognitive dissonance among users, or the cost of sharing racist messages increases the latency of the decision to bare the (social) cost of sharing racist content. The results are in line with the expectations from Hypothesis 1.

There are other particularities about the network worth mentioning. The pro-government community *was not* eager to retweet messages from their own authorities. Results show that high-degree authorities (above the log-median in-degree) were retweeted 20% slower than low-degree authors (below the log-median in-degree). However, high-degree hubs retweeted posts 9% faster than low-degree hubs. The differences in behavior shows that the "soldiers" of the network–i.e. low-degree yet high-activity users–were not being activated by the content posted by the authorities, but rather by content created by other, less prominent, users.

Not all users will engage with racist content similarly. Users

[§§§] Twitter rules includes the following: "You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." A violation to this rule can result in the elimination of the tweet.

[¶¶¶] To reduce chance of losing deleted tweets, I ran Twitter's backward search multiple times during the period analyzed.

[17] Using time-to-retweet has additional empirical benefits. Given the network structure of Twitter, a racists tweet might not reach all users. The exposure to a message is closely linked to the number of followers each users has. Thus, depending on the source, there are users that will not be exposed at all, and therefore not have the opportunity to engage with racist tweets. By looking at time-to-retweet we can see the differences in reactions, independent of exposure.

[18] Verified users in the Ecuadorian network are also high-degree users. This is to be expected.

[19] Proportional Hazard Cox models explain survival rates, which in our case explain *time-to-retweet*.

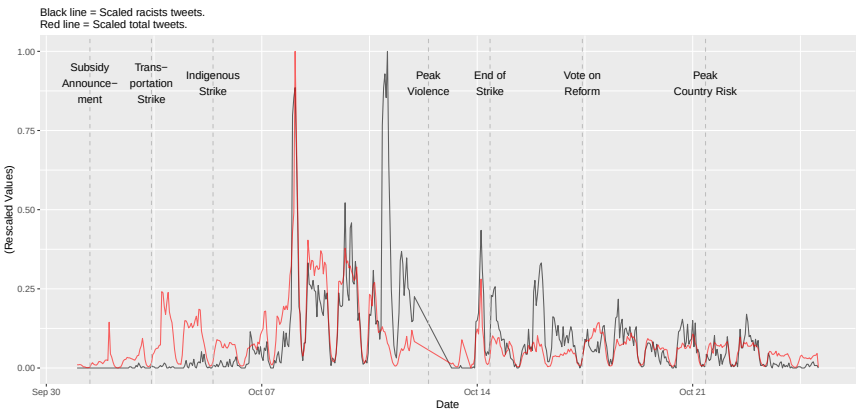


Fig. 2. Timeline of #ParoEcuador activity, between October 1 and October 24 2019.

Table 1. Racist content and time-to-retweet in the Ecuadorian pro-government community

	Model 1	Model 2	Model 3
Racist Attack	-0.148***	-0.189***	-0.128***
	(0.012)	(0.016)	(0.014)
Auth Bot Control	-0.022***	-0.022***	-0.022***
	(0.008)	(0.008)	(0.008)
Hub Bot Control	-0.292***	-0.292***	-0.292***
	(0.028)	(0.028)	(0.028)
High Degree Auth (Dummy)	-0.216***	-0.216***	-0.215***
	(0.004)	(0.004)	(0.004)
High Degree Hub (Dummy)	0.085***	0.084***	0.085***
	(0.003)	(0.003)	(0.003)
Friends Auth (ln)	0.149***	0.149***	0.149***
	(0.008)	(0.008)	(0.008)
Followers Auth (ln)	0.095***	0.095***	0.095***
	(0.002)	(0.002)	(0.002)
Friends Hub (ln)	0.241***	0.241***	0.241***
	(0.020)	(0.020)	(0.020)
Followers Hub (ln)	0.065***	0.065***	0.065***
	(0.015)	(0.015)	(0.015)
Racist Attack * High Degree Auth		0.086***	
		(0.023)	
Racist Attack * High Degree Hub			-0.063**
			(0.025)
N	465,511	465,511	465,511

Note: Hazard estimates of time-to-retweet. Positive numbers indicate positive increases in hazard rate and shorter time to retweet. Standard errors are reported in parentheses, with confidence levels reported as follows: ***p < .01; **p < .05; *p < .1.

across different network topographies can have different reactions to the same content.[20] I explore two groups of interest: high-degree and low-degree users. To estimate the different reactions I interact high- and low-degree users and racist content, and plot the survival probability curves from the interaction in Figure 3. From Model 1 in Table 1 we know that racist tweets in general lengthened time-to-retweet among pro-government users. Indeed, both low-degree and high-degree users were slow to retweet messages that featured racist content.

However, the interaction between high-degree users and racist content (Figure 3, left panel) shows that the slowing effect (cognitive dissonance) of racist content is more pronounced in low-degree users (hubs) than in high-degree users. The penalty of racist content on time-to-retweet remains in high- and low-degree users, but the former penalize racist content substantively less. For a comparison, on average, a high-degree user will retweet a racist message as fast as a low-degree user will retweet a non-racist message. The interaction between high-degree *authors* and racist content (Figure 3, right panel) show a similar trend. On average, users in the pro-government network will have longer time-to-retweet for racist content. However, racist content produced by high-degree authors create more cognitive dissonance on users, than similar content produced by low-degree users. Thus, we find that racist content is mobilizing high-degree users faster than low-degree users. Furthermore, when it is high-degree author positioning these frames, racist content produces greater cognitive dissonance than what you would expect from racist content more generally.

To summarize, as anticipated by Hypothesis 1, racist content had higher time-to-retweet than non-racist content. The larger latency in time-to-retweet suggests that, in general, racist content induced higher cognitive dissonance. Furthermore, the results suggest that public figures avoid producing racist content and avoid reproducing it. The social cost of overt racism is a likely deterrent for that behavior, especially damaging to the image of more prominent political personalities. Alternatively, more public figures might also be less racist. While this possibility cannot be discarded, it seems less plausible in light of the racist structure in which this network is embedded. Additionally, I show that high-degree users, users closer to power and more immediately affected by the out-group, have shorter time-to-retweet of racist content than low-degree users. In other words, within the pro-government community, it is those closer to the authorities, rather than the "soldiers," that find racist frames more cognitively congruent when contenting with an out-group.

Threats to the In-Group and Racist Contents

The main argument of this paper is that racism is activated when the status of the in-group is threatened (H2). To provide empirical evidence, I look at two events where the indígena community (out-group) either threatened the status of the gov-

[20] Calvo and Aruguete (2018) describe this as the 'composition effect.'

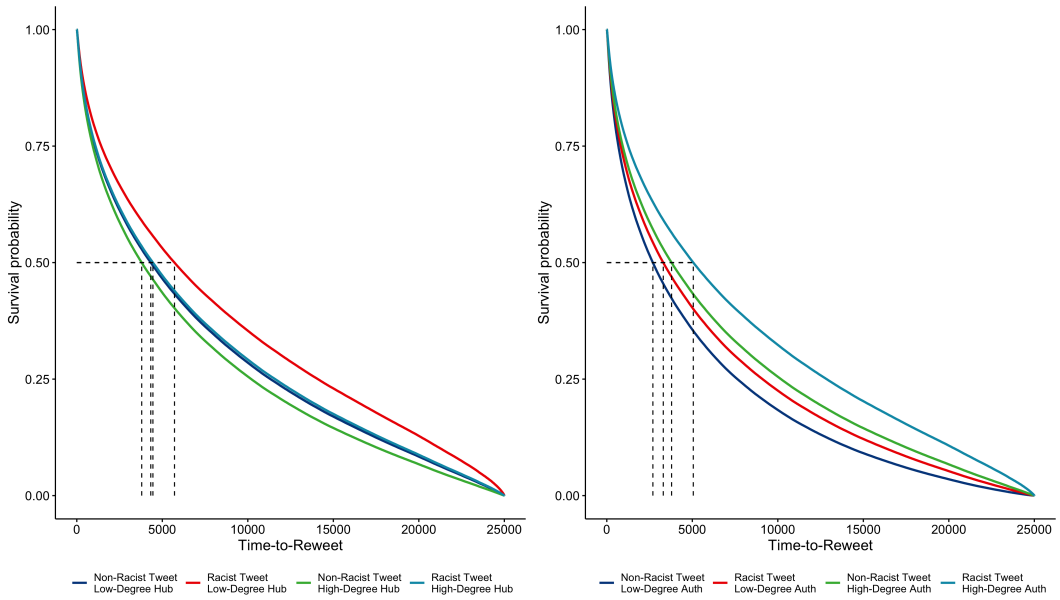


Fig. 3. Survival probability of Proportional Hazard Cox modes from Table 1, Models 2 and 3, with time-to-retweet as the dependent variable. On the left panel: survival probability of the interaction between racist tweets and high- and low-degree *hub* (user trait). On the right panel: survival probability of the interaction between racist tweets and high- and low-degree *authors* (author trait). All other variables kept at their means.

ernment, either by calling forces to challenge the government, or when the government acquiesced to the indígena community's demands. Thus, I test my hypothesis by examining the moment when 1) Jaime Vargas, the president of the CONAIE, called the policy and military forces to disobey government orders and join the protest, and 2) when the Moreno government announced the end of the strike after agreeing to the demands of the indígena leadership.

Both episodes were highly salient events. High-salience events focus the attention of the public and redefines the situation (Pride 1995). Lin et al. (2013) argue that highly salient media events "create a social condition call 'shared attention,' [...] where the larger potential audiences, altered norms, and the high level of *shared understanding* can all contribute to shifts in the ways in which users [...] attend to content." (emphasis mine). Lin et al. (2013) go on to explain that in media events, users can converse without needing to explain the context, allowing for the creation of a more disciplined message. Media event that result in adjudication, for example when Moreno's government negotiated the end of the strike with indígena leaders, should be particularly triggering of racist content.

Calls to Arms and Acquiescing

Historically, the indígena uprisings have had three political outcomes: subjugation through violence, policy reform and expansion of rights (e.g. Constitutional recognition of Ecuador as a plurinational state), and coups. The 2019 protests in Ecuador began as a series of demands (i.e. policy reform) made by the indígena community in light of the elimination of gas subsidies. Indígenas began marching from their communities, often located in rural Ecuador or in the *Amazonía*, and headed, as is customary, to the capital, Quito, where they would occupy the main political markers and negotiate with the government.

On October 10, the indígenas assembled in a large forum in Quito where eight police officers, among them a colonel, were detained for more than two hours, as well as various journalists and their crews.[21] Jaime Vargas, the president of the CONAIE, was speaking to the crowd on live television (the detained journalist and their crews came in handy) with the detained police officers in the backdrop. At 11:51 AM, local time, Vargas asks the detained colonel, the Police force, and the military, to join the protests and disobey the orders of the government.

The moment Vargas calls for the sedition of the police and armed forces creates a reaction in the pro-government community. Given recent Ecuadorian political history,[22] the call from Vargas is not an empty threat, but rather a clear challenge to power and the status of the government. Through Vargas' call, the out-group conspicuously challenges the in-group. We would expect that this action, a threat to the in-group, will trigger the shared identity of the pro-government community, in particular the racial identity.

Similarly, on October 13, the high-ranking officials from the Moreno government sat down with indígena leaders to negotiate the end of the strike. The indígena leaders demanded, among other things, Moreno to rescind the order that prompted the mobilizations in the first place. After three hours of negotiations, the meeting was broadcasted live, as each part made public statements. At 9:45 PM, local time, a United Nations mediator announced the conditions under which the strike would be over. Lenín Moreno agreed to reinstate the gas subsidies and the indígena leaders agreed to suspend the mobilization. In the eyes of pro-government users, the indígenas had won. The adjudication of the strike was

[21] The details of the detainment are unclear. The indígena community blamed the Police and the government for the deaths of various indígena protesters. This was the main reason to hold the police officers. The government said the police officers were kidnapped, while the indígena leaders, mainly Jaime Vargas, argued that the officers and the press could leave at any time.

[22] For example, in the 2000 Ecuadorian coup that unseated President Jamil Mahuad, then President of the CONAIE, Antonio Vargas, was joined by the highest ranking Army General and the President of the Supreme Court in a triumvirate that briefly held power.

the materialization of the challenges to the in-group status. Like in the previous event, the immediate actions that led to the end of the strike were broadcasted and garnered collective attention. As I will show, in both events pro-government users reacted to these threats to the status of the in-group by engaging with racist frames.

To determine the effect of these events, I follow Calvo et al. (2020) and use an interrupted time series analysis, a variety of regression discontinuity designs (RDD) in which the running variable is time (Morgan and Winship 2015; Mummolo 2018). Twitter data is ideal for this approach because of the granularity and high-frequency of tweets. Our primary parameter of interest is the change in users' latency (i.e. time-to-retweet) when engaging with racist content upon Vargas' call to the Police and the Army and after the announcement to end the strike. The time of each action is the cut-off of the regression model.

Regression discontinuity models assume that effects are continuous at the cut-off (De la Cuesta and Imai 2016). When dealing with time as a running variable, the continuity assumption requires that no omitted variable that systematically affects the outcome (i.e. racist content and time-to-retweet) also changes upon the occurrence of the event (i.e. Vargas' call to the Police and Army or the announcement of the end of the strike). Given that we have the precise moment when the two events were broadcasted live, and estimating the models within a six hours window around cut-off, it is reasonable to assume that this assumption holds. The granularity of the data, together with the precise measurement of the event, makes the identification strategy highly plausible. Appendix B provides a set of tests to verify the continuity assumption, including placebo checks with the running variable. Overall, the results ensure the internal validity of the RD design.

To estimate the models I follow Gelman and Imbens (2019) and use a non-parametric local linear regression (LLR) to approximate the treatment effect at the cut-off points. I employ a local polynomial of order $p = 1$, which gives the standard local linear regression discontinuity point estimator, and fit two separate regression functions above and below the cutoff event. The treatment effect is the difference in the limits of the cutoff. I use a data-drive mean squared error search to select optimal bandwidths and a triangular kernel that assigns linear down-weighting to each observation. In Tables C.1 and C.2 (see Appendix C) I report the robust bias-corrected treatment effects and 95% confidence intervals as described by Calonico et al. (2014).

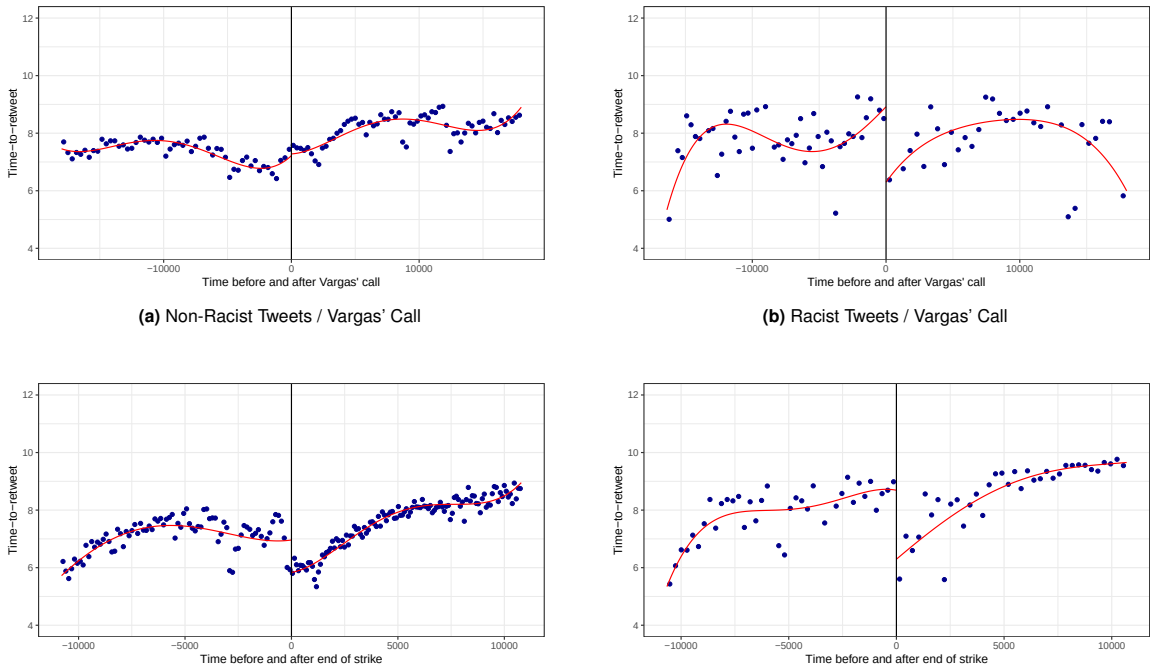
Panels (a) and (b) in Figure 4 provide evidence of the effect Vargas' speech on pro-government users. For both figures the vertical axis reports time-to-retweet and it is interpreted as usual: lower values mean less time-to-retweet and user latency. The sample is divided into racist and non-racist tweets. The horizontal axis has a range of twelve hours, six hours before and after the event. I use a LOESS smoother to fit the underlying regression function separately before and after the event. The discontinuity shows that at the time of Vargas' speech, there is no change in latency for users sharing non-racist content, but there is a statistically and substantively significant difference in latency for users sharing racist message ($p \leq 0.05$; see Appendix C for full model results). Immediately after Vargas' call, racist content increases engagement and reduces latency among pro-government users. Notice that before the cutoff, racist content is spread, on average, slower than non-racist content. However, at the moment of Vargas' speech, users engage with racist content faster than with non-racist content. The reaction to the threat to the status of the in-group made racist frames cognitively congruent to pro-government users.

A similar treatment effect can be observed when the end of the strike was announced (see panels (c) and (d) in Figure 4). In this case, the discontinuity shows that the end of the strike decreased the latency for both non-racist and racist content ($p \leq 0.05$). Nonetheless, the treatment effect for racist tweets (-2.17) is larger than the treatment effect for non-racist tweets (-0.66). Analogous to Vargas' call, racist content went from having longer time-to-retweet than non-racist content, to roughly the same when the end of the strike announced. In other words, both types of frames created similar cognitive congruence in pro-government users after the strike adjudication.

The results from the previous section show that in our pro-government community, high-degree users engage significantly faster with racist frames than low-degree users. To estimate how high- and low-degree users react to the challenges to the in-group, I further divide the sample into high-degree and low-degree users, and estimate the changes in time-to-retweet at the cutoff for racist content. Again, panels (a) and (b) in Figure 5 show treatment effect of Vargas' call. The treatment effect is only statistically significant for high-degree users ($p \leq 0.05$; see Appendix C for full model results). While low-degree users also engage faster with racist content, the difference is neither statistically significant nor evident from the graph. On the contrary, the end of the strike produced a homogenous effect on the reaction to racist content from low- and high-degree users (see panels (c) and (d) in Figure 5). Not only is the treatment effect statistically significant for both types of users ($p \leq 0.05$), but the magnitude of the effect (log time-to-retweet) is similar: -2.6 for high-degree users and -1.9 for low-degree users.

To further understand the reaction to racist content, I look at changes in the composition of the pro-government community before and after both events. Following Lin et al. (2014), Figure 6 plots the average degree of users (engaging with racist content) against its concentration as measured by the Gini coefficient. The Gini coefficient (y-axis) measures the level of the pro-government community's degree distribution. A lower Gini coefficient indicates a more equal distribution. According to Lin et al. (2014), horizontal movements indicate "rising tides" or system-level changes, increases in connectivity without changes in concentration. Contrary, vertical movement indicates "rising stars" or individual-level changes, increases in concentration without changes in connectivity.

Panel (a) in Figure 6 corresponding to the before and after Vargas' calling, shows little movement either horizontally or vertically. This means that there were almost no aggregate changes at the system level (i.e. no "rising tides"), and not changes at the individual level (i.e. no "rising stars"). This would explain why, despite the changes in behavior–decreases in the overall latency–the differences between the high- and low-degree users remained. Furthermore, it hints at the possibility that the event itself was not a large-enough (or concrete-enough) threat to the pro-government community to rally users around a racial identity. A different picture is shown of the pro-government community before and after the end



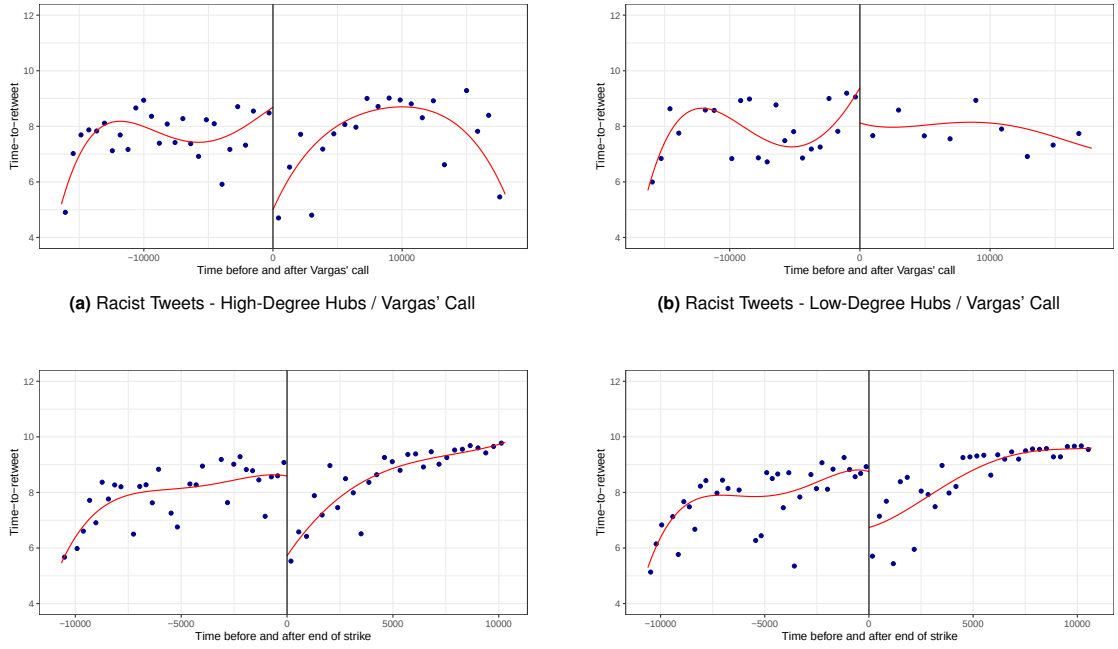
(a) Non-Racist Tweets / Vargas' Call

(b) Racist Tweets / Vargas' Call

(c) Non-Racist Tweets / End of Strike

(d) Racist Tweets / End of Strike

Fig. 4. Time-to-Retweet during the Ecuadorian protests. Top panels: Centering on October 10, 2019, at 11:51 AM, local time, when Jaime Vargas, the president of the CONAIE, asks the Police force and the military to join the protests and disobey the orders of the government. Bottom panels: Centering on October 11, 2019, at 9:45 PM, local time, when the United Nations mediator announced the terms agreed upon for the end of the strike.



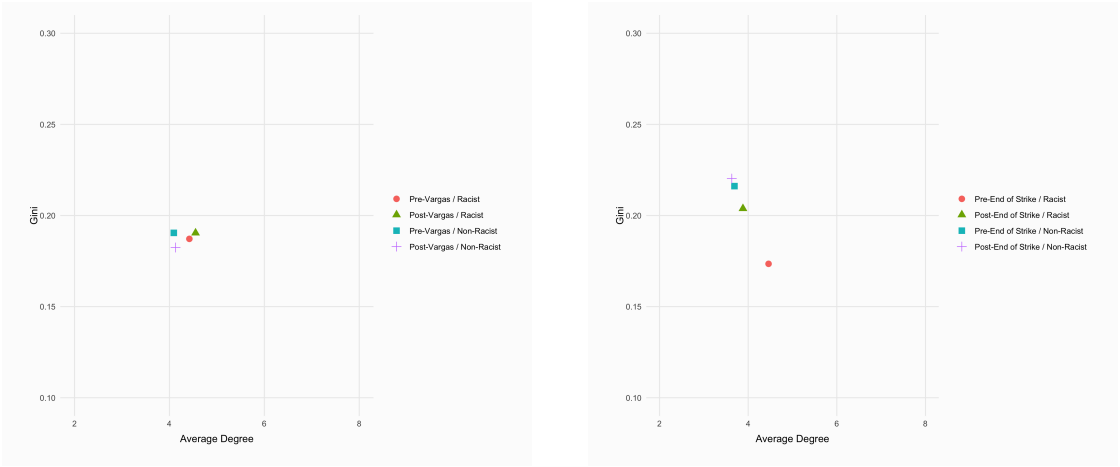
(a) Racist Tweets - High-Degree Hubs / Vargas' Call

(b) Racist Tweets - Low-Degree Hubs / Vargas' Call

(c) Racist Tweets - High-Degree Hubs / End of Strike

(d) Racist Tweets - Low-Degree Hubs / End of Strike

Fig. 5. Time-to-Retweet during the Ecuadorian protests. Top panels: Centering on October 10, 2019, at 11:51 AM, local time, when Jaime Vargas, the president of the CONAIE, asks the Police force and the military to join the protests and disobey the orders of the government. Bottom panels: Centering on October 11, 2019, at 9:45 PM, local time, when the United Nations mediator announced the terms agreed upon for the end of the strike.



(a) Hubs / Vargas' Call

(b) Hubs / End of Strike

Fig. 6. (a) shows average degree and Gini coefficient for before and after Vargas' call. (b) shows average degree and Gini coefficient for before and after the end of strike announcement. The Gini coefficient (y-axis) measures the level of the pro-government community's degree distribution. A lower Gini coefficient indicates a more equal distribution. For all graphs, the sample is limited to users engaging with racist content.

of the strike (panel (b) in Figure 6). The upwards vertical shift for users engaging with racist content suggests the hub activity of the network became more centralized. This would suggest that racist content became more cognitive congruent to a smaller, yet more active, set of users.

Overall, I find that users of the in-group, in our case to the pro-government community, engaged with racist frames when the status of the in-group was threatened. The analysis of two high-salience events shows that cognitively congruent reactions from pro-government users to racist content. However, while more concrete threats, like the announcement of the end of the strike, had a homogenous effects across the topography of the community, other events, like Vargas' call to the police and Army to join the strike, did not. I also find, absent these threats, racist content creates cognitive dissonance to users, even though, for this particular network, high-degree users engage faster with racist tweets than low-degree users.

Conclusions

This research explores the spread of racist content in social media. Taking cues from the political psychology literature on out-group hate and the communications literature on online behavior, I expect racist frames to activate as a reaction to threats to the in-group. Given the social cost of racism and the real consequence that an online action can have on a user, there will be variation on who reproduces racist messages, and when are these frames activated.

This paper contributes on various fronts. The analysis of the Ecuadorian protests of 2019 shows that threats by the indígena community (out-group) to the status of the government (in-group) make racist frames cognitively congruent to pro-government users. While racist content is present in our Twitter network throughout the development of the protest, it required concrete threats, like the acquiescence of the government to the indígena demands, for userss to actively engage with racist frames.

Methodologically, I provide an easy-to-implement process to identify racist tweets. I use the machine-learning algorithm *Perspective* to detect identity attacks in our corpus. I then combine these results with a list of terms that serve as markers for the contextual forms of racism found in Ecuador. Unlike alternative semi-supervised machine-learning approaches that require hard-to-come-by tagged corpora, *Perspective* is pre-trained in various languages. By providing contextual information of the racist discursive forms I am able to effectively detect racism avoiding, primarily, a large number of false positives. The method has some limitations: it is not able to identify more subtle forms of racism (and more reflective of the extent of racist ideology) and it over-represents racism in communities where racism is unlikely to be produced. I believe that these are two areas where this process can improve in future research.

Polarized environments such as those found in Ecuador help detect the behavior of racist content, a problem that is pervasive to most political contexts. I identify some mechanisms that explain the proliferation of racist discourses in social media, important to know how to eventually stop them. Aggressiveness and toxicity are not endemic to the community representing the dominant group in society. In our network I find high levels of both in our indígena community as well. However, aggressiveness and toxicity towards power is less problematic than the other way around. In the case of Ecuador, the racist ideology excluded the indígena community from the participating in the political process and the construction of the national identity. Our network analysis reveals how, in this context, racism is propagated, a mechanism that has yet to be systematically studied in social media.

References

Amira, K., Wright, J. C., and Goya-Tocchetto, D. (2019). In-group love versus out-group hate: Which is more important to partisans and when? *Political Behavior*, pages 1–22.

Aruguete, N. and Calvo, E. (2018). Time to# protest: Selective exposure, cascading activation, and framing in social media. *Journal of communication*, 68(3):480–502.

Back, L. (2002). Aryans reading adorno: cyber-culture and twenty-first century racism. *Ethnic and racial studies*, 25(4):628–651.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., and Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542.

Becker, M. (2010). *Pachakutik: Indigenous movements and electoral politics in Ecuador*. Rowman & Littlefield Publishers.

Bobo, L., Kluegel, J. R., and Smith, R. A. (1997). Laissez-faire racism: The crystallization of a kinder, gentler, antiblack ideology. *Racial attitudes in the 1990s: Continuity and change*, 15:23–25.

Bonilla-Silva, E. (2015). The structure of racism in color-blind,'post-racial' america.

Bretón, V. and Pascual, F. G. (2003). *Estado, etnicidad y movimientos sociales en América Latina: Ecuador en crisis*, volume 27. Icaria Editorial.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, 55(3):429–444.

Brock, A. (2009). Life on the wire: Deconstructing race on the internet. *Information, Communication & Society*, 12(3):344–363.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cervone, E. and Rivera, R. V. (1999). *Ecuador racista: imágenes e identidades*. FLACSO, Sede Ecuador.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22.

Chaudhry, I. and Gruzd, A. (2020). Expressing and challenging racist discourse on facebook: How social media weaken the "spiral of silence" theory. *Policy & Internet*, 12(1):88–108.

Christopherson, K. M. (2007). The positive and negative implications of anonymity in internet social interactions:"on the internet, nobody knows you're a dog". *Computers in Human Behavior*, 23(6):3038–3056.

Colloredo-Mansfeld, R. (1998). 'dirty indians', radical indígenas, and the political economy of social difference in modern ecuador. *Bulletin of Latin American Research*, 17(2):185–205.

Csardi, G., Nepusz, T., et al. (2006). The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9.

Daniels, J. (2013). Race and racism in internet studies: A review and critique. *New Media & Society*, 15(5):695–719.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

De la Torre Espinosa, C. (1996). El racismo en el ecuador: experiencia de los indios de clase media. In *El racismo en el Ecuador: Experiencia de los indios de clase media*, pages 111–111.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. *arXiv preprint arXiv:1804.04649*.

Eschmann, R. (2020). Unmasking racism: Students of color and expressions of racism in online spaces. *Social Problems*, 67(3):418–436.

Fox, J., Cruz, C., and Lee, J. Y. (2015). Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior*, 52:436–442.

Gelman, A. and Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, 37(3):447–456.

Groenendyk, E. W. and Banks, A. J. (2014). Emotional rescue: How affect helps partisans overcome collective action problems. *Political Psychology*, 35(3):359–378.

Guerrero, A. (1997). The construction of a ventriloquist's image: liberal discourse and the 'miserable indian race'in late 19th-century ecuador. *Journal of Latin American Studies*, 29(3):555–590.

Hall, G. and Patrinos, H. A. (2004). *Indigenous peoples, poverty and human development in Latin America: 1994-2004*. The World Bank.

Himelboim, I., Smith, M., and Shneiderman, B. (2013). Tweeting apart: Applying network analysis to detect selective exposure clusters in twitter. *Communication methods and measures*, 7(3-4):195–223.

Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

Huddy, L. (2001). From social to political identity: A critical examination of social identity theory. *Political psychology*, 22(1):127–156.

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

Kaakinen, M., Sirola, A., Savolainen, I., and Oksanen, A. (2020). Shared identity and shared information in social media: development and validation of the identity bubble reinforcement scale. *Media Psychology*, 23(1):25–51.

Keipi, T., Näsi, M., Oksanen, A., and Räsänen, P. (2016). *Online hate and harmful content: Cross-national perspectives*. Taylor & Francis.

Lapidot-Lefler, N. and Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, 28(2):434–443.

Martínez-Echazábal, L. (1998). Mestizaje and the discourse of national/cultural identity in latin america, 1845-1959. *Latin American Perspectives*, 25(3):21–42.

Mason, L. (2016). A cross-cutting calm: How social sorting drives affective polarization. *Public Opinion Quarterly*, 80(S1):351–377.

Morgan, S. L. and Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Mummolo, J. (2018). Modern police tactics, police-citizen interactions, and the prospects for reform. *The Journal of Politics*, 80(1):1–15.

Nakamura, L. (2008). *Digitizing race: Visual cultures of the Internet*, volume 23. U of Minnesota Press.

Neumayer, C., Rossi, L., and Karlsson, B. (2016). Contested hashtags: blockupy frankfurt in social media. *International Journal of Communication*, 10:22.

Omi, M. and Winant, H. (2014). *Racial formation in the United States*. Routledge.

Pride, R. A. (1995). How activists and media frame social problems: Critical events versus performance trends for schools. *Political Communication*, 12(1):5–26.

Roitman, K. (2009). *Race, ethnicity, and power in Ecuador: The manipulation of mestizaje*. FirstForumPress.

Roitman, K. and Oviedo, A. (2017). Mestizo racism in ecuador. *Ethnic and racial studies*, 40(15):2768–2786.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media*, pages 1–10.

Sue, C. A. and Golash-Boza, T. (2013). 'it was only a joke': How racial humour fuels colour-blind ideologies in mexico and peru. *Ethnic and Racial Studies*, 36(10):1582–1598.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. Cup Archive.

Timoneda, J. C. (2018). Where in the world is my tweet: Detecting irregular removal patterns on twitter. *PloS one*, 13(9):e0203104.

Van Cott, D. L. (2008). *Radical democracy in the Andes*. Cambridge University Press Cambridge.

Van Dijk, T. A. (1993). *Elite discourse and racism*, volume 6. Sage.

van Dijk, T. A. (2005). *Racism and discourse in Spain and Latin America*, volume 14. John Benjamins Publishing Company Amsterdam.

Van Dijk, T. A. et al. (2000). New (s) racism: A discourse analytical approach. *Ethnic minorities and the media*, 37:33–49.

Vélez, F. R. (2000). Los indigenismos en ecuador: de paternalismos y otras representaciones. *Cuadernos de antropología: Revista Digital del Laboratorio de Etnología" María Eugenia Bozzoli Vargas"*, 11(1):97–108.

Vinueza, J. A. (1999). Regionalismo y movimiento indígena en el ecuador: un reto a la política de la diferencia. *Boletín de antropología americana*, (35):113–124.

Whitten Jr, N. E. (2003). Symbolic inversion, the topology of el mestizaje, and the spaces of las razas in ecuador. *Journal of Latin American Anthropology*, 8(1):52–85.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.