On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook

Filde N. Ribeiro*, UFOP/UFMG, Brazil  
filipe.ribeiro@ufop.edu.br  
  
Koustuv Saha*  
Georgia Tech, US  
koustuv.saha@gatech.edu  
  
Mahmoudreza Babaei  
MPI-SWS, Germany  
babaei@mpi-sws.org  
  
Lucas Henrique  
UFMG, Brazil  
lhenriquecl@dcc.ufmg.br  
  
Johnnatan Messias  
MPI-SWS, Germany  
johnmne@mpi-sws.org  
  
Fabricio Benevenuto  
UFMG, Brazil  
fabricio@dcc.ufmg.br  
  
Oana Goga  
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, France  
oana.goga@univ-grenoble-alpes.fr  
  
Krishna P. Gummadi  
MPI-SWS, Germany  
gummadi@mpi-sws.org  
  
Elissa M. Redmiles  
University of Maryland, US  
eredmiles@cs.umd.edu

ABSTRACT
Targeted advertising is meant to improve the efficiency of matching advertisers to their customers. However, targeted advertising can also be abused by malicious advertisers to efficiently reach people susceptible to false stories, stoke grievances, and incite social conflict. Since targeted ads are not seen by non-targeted and non-vulnerable people, malicious ads are likely to go unreported and their effects undetected. This work examines a specific case of malicious advertising, exploring the extent to which political ads from the Russian Intelligence Research Agency (IRA) run prior to 2016 U.S. elections exploited Facebook’s targeted advertising infrastructure to efficiently target ads on divisive or polarizing topics (e.g., immigration, race-based policing) at vulnerable subpopulations. In particular, we do the following: (a) We conduct U.S. census-representative surveys to characterize how users with different political ideologies report, approve, and perceive truth in the content of the IRA ads. Our surveys show that many ads are “divisive”: they elicit very different reactions from people belonging to different socially salient groups. (b) We characterize how these divisive ads are targeted to subpopulations that feel particularly aggrieved by the status quo. Our findings support existing calls for greater transparency of content and targeting of political ads. (c) We particularly focus on how the Facebook ad API facilitates such targeting. We show how the enormous amount of personal data Facebook aggregates about users and makes available to advertisers enables such malicious targeting.

1 INTRODUCTION
Online targeted advertising refers to the ability of an advertiser to select an audience for their ads. Such advertising constitutes the primary source of revenue for many online sites including most social media websites such as Facebook, Twitter, YouTube, and Pinterest. Consequently, these websites accumulate detailed demographic, behavioral and interest profiles of their users enabling advertisers to “microtarget”, i.e., choose small (tens or hundreds to thousands) of users with very specific attributes like people living in a zipcode that read New York Times or Breitbart. Beyond raising numerous privacy concerns where ads announcing housing or job opportunities are targeted to exclude people belonging to certain races or gender [4, 8, 10, 23].

In this paper, we analyze the potential for a new form of abuse on targeted advertising platforms namely, socially divisive advertising, where malicious advertisers incite social conflict by publishing ads on divisive societal issues of the day (e.g., immigration and racialbias in policing in the lead up to 2016 US presidential elections). Specifically, we focus on how ad targeting on social media sites such as Facebook can be leveraged to selectively target groups on different sides of a divisive issue with (potentially false) messages...
that are deliberately crafted to stoke their grievances and thereby, worsen social discord. We also investigate whether targeted ad platforms allow such malicious campaigns to be carried out in stealth, by excluding people who are likely to report (i.e., alert site administrators or media watchdog groups about) such ads.

Our study is based on an in-depth analysis of a publicly released dataset of Facebook ads run by a Russian agency called Internet Research Agency (IRA) before and during the American Election on the year of 2016. Our analysis is centered around three high-level research questions:

RQ 1: How divisive is the content of the IRA ads? We quantify the divisiveness of an ad by analyzing the differences in reactions of people with different ideological persuasions to the ad. Specifically, using US census-representative surveys, we look at how conservative- and liberal-minded people differ in (a) how likely they are to report the ad, (b) how strongly they approve or disapprove the ad’s content, and (c) how they perceive truthhood (or falsehood) in ad’s claims. Our analysis shows that IRA ads elicit starkly different and polarizing responses from people with different ideological persuasions.

RQ 2: How effectively done was the targeting of the socially divisive ads? We find that the “Click Through Rate” (CTR), a traditional measure of effectiveness of targeting, of the IRA ads are an order of magnitude (10 times) higher than that of typical Facebook ads. The high CTR suggests that the ads have been targeted very efficiently. A deeper analysis of the demographic biases in the targeted audience reveals that the ads have been exploited at people who are more likely to approve the content and perceive fewer false claims, and are less likely to report.

RQ 3: What features of Facebook’s ad API were leveraged in targeting the ads? We also analyze the construction or specification of “targeting formulae” for the ads, i.e., the combination of Facebook user attributes that are used when selecting the audience for the ads. We find widespread use of interest attributes such as “Black Consciousness movement” and “Chicano movement” that are mostly shared by people from specific demographic groups such as African-Americans and Mexican-Americans. We show how Facebook ad API’s suggestion feature may be exploited by the advertisers to find interest attributes that correlate very strongly to specific social demographic groups.

1.1 Related Work

Prior work has highlighted several forms of abuses of targeted advertising in Facebook, such as for inappropriately exposing the private information of users to advertisers [24], and for allowing discriminatory advertising (e.g., to exclude users belonging to a certain race or gender from receiving their ads) [23]. Our effort highlights a new and different form of potential abuse of these targeted advertising platforms in creating a social discord.

A rich body of prior work have focused on understanding filter bubbles, echo chambers, polarization, and ideological discourse in social media as an emergent phenomenon [7, 9, 12, 14, 15, 19, 22]. We provide a complementary perspective on the topic by examining how echo chambers and polarization can be engineered on social media through targeted advertising. A recent work conducted a detailed study about Facebook Ads environment by analyzing thousands of ads collected through a browser plugin[2]. More closely related to our work, Kim et al. gathered Facebook ads from individuals and analyzed who are behind divisive ad campaigns, reporting suspicious foreign entities [16]. Differently, we focus on understanding the disruptive ability of microtargeting for providing divisive political ad campaigns.

Finally, our effort is complementary to prior work that attempts to understand the abuse of social media by misinformation campaigns, especially along political elections [18, 25]. Our work provides a better comprehension about a key dissemination mechanism of fake news stories, highlighting how advertising platforms allow injection of misinformation in social systems and choose vulnerable people as the target.

2 RUSSIA-LINKED FACEBOOK ADS DATASET

On May 10th, 2018 the Democrats Permanent Select Committee on Intelligence released a dataset containing 3,517 Facebook advertisements from 2015, 2016, and 2017 that are linked to a Russian propaganda group: Internet Research Agency (IRA). Each ad is composed of an image and text (Figure 1 shows an example). Additionally, each ad contains a landing page, which is a link to the host of the ad, as well as an ad ID; an ad targeting formula, which is a combination of demographic, behavioral and user interest aspects used to target Facebook users; the cost for running the ad in Russia Rubles; the number of impressions, which is the number of users who spent some time observing the ad; the number of clicks received by the ad; and, finally, the ad creation and end dates. This section provides an overview of these ads.

The ads in the dataset were run between June 2015 and August 2017. From the 3,517 advertisements, we found that 617 (17.5%) were created in 2015, 1,867 (53.1%) in 2016, and 1,033 (29.4%) in 2017. Figure 2 shows the distribution of these ads over time in terms of the number of ads created per month, cost to run the ads, and

---

1. www.wsj.com/articles/you-cant-buy-the-presidency-for-100-000-1508104629
3. democrats-intelligence.house.gov/facebook-ads/social-media-advertisements.htm
4. We converted currency of the costs to USD as of May 15th, £1 USD = 61.33 RUB.
We first explore the ad landing pages: the number of impressions, clicks, and amount spent to advertise. The most expensive ad cost 5,307 USD. The highest number of impressions generated was 1,335,000 and the maximum number of clicks was 73,060.

Nearly 25% of the landing pages spent more than 100 dollars, 26.8% of the pages received more than 1,000 clicks, and around 36.1% had more than 10,000 impressions. On the other hand, more than 25% of the ads had no impressions, clicks, and cost, suggesting these ads were not launched or ran for a very short period of time.

An average ad cost 34.5 USD, was seen by 11,536 users, and received 1,062 clicks. The average value is increased to 38 USD for cost, 16,482 for impressions, and 1,521 for the number of clicks if we exclude those ads that appeared not to have been run. The Pearson’s correlation coefficient among cost, impressions, and clicks is very high, particularly between impressions and clicks (0.89). We also noted that this dataset is quite skewed, as 10% of the ads accumulate 85.18% of the total cost, 71.93% of the total number of impressions, 69.47% of the total number of clicks.

However, there were notable exceptions to this correlation: higher investment (cost) did not always lead to higher return (e.g., impressions, clicks). Table 1 shows the most popular landing pages per impressions, clicks, and cost of the ads. For example, fb.com/ blackmattersus/.mvmt/ received the largest number of impressions (5,817,734), corresponding alone to 14.3% of impressions obtained by all ads, but cost only 6.5% of the total cost of all ads in the dataset.

Finally, we compute the click-through rate (CTR) of these ads, which is a typical metric to measure the effectiveness of an ad. It is computed as a ratio between the number of clicks and the number of impressions received by an ad. Figure 4 (right) shows the cumulative distribution function of the CTR of the ads, excluding those with 0 values for clicks, impressions, and cost. The median CTR is 10.8% and 75% of the ads have a CTR higher than 5.6. The average CTR is 10.8%. These are incredibly high values for CTR. As a comparison, WordStream released a report as of April 2018 which shows the average CTR for Facebook ads across all industries is 0.9%. As an example, Retail is 1.6%, Fitness is 1%, Health care 0.8%, and Finance is 0.56%. This means that these political ads have a CTR that is about an order of magnitude higher than a typical Facebook ad.

2.3 High Impact Ads

Our analysis reveals that only a few ads are responsible for most of the cost, impressions, and clicks. Considering this, we defined a set of high impact ads as the union of the top 10% ads in terms of cost, impressions, clicks, and CTR. We obtained 905 high impact ads, corresponding to 27.7% of the entire dataset. These ads account together to 83.9% of the total number of impressions, 81.8% of clicks, 88.5% of the cost, and 46.9% of the CTR. For the purposes of our study, where we require manual inspection of the ads (to identify their targets and to run surveys), our ensuing analyses concern those high impact ads run before the 2016 U.S. elections: 485 ads.
We find that the ad campaigns were intensified near to the U.S. election period. Among our main findings, we show that the typical CTR for these ads is an order of magnitude higher than typical election period. Among our main findings, we show that the typical CTR for these ads is an order of magnitude higher than typical ads which were run before the elections. Each survey showed ten ads followed with different political viewpoints – we conducted three online surveys on a U.S. census-representative sample (n=2,886). We used each survey to measure one of three axes along which ads could potentially be divisive: 1) reporting: whether respondents would report the ads, and why, 2) approval and disapproval: whether they approve or disapprove the content of the ad, and 3) false claims: if they are able to identify any false claims in the content of the ad.

Our surveys considered only those 485 ads which were run before the elections. Each survey showed ten ads followed by demographic questions. More detail on the specific questions used to assess each axis is provided in the corresponding axis subsections that follow. The survey questions were pre-tested using cognitive interviews and all survey questions included a “I don’t know” or “Prefer not to respond” answer choice to ensure internal measurement validity [6]. To obtain a demographically representative sample, and ensure that we captured a wide variety of American perceptions, we deployed the surveys using the Survey Sampling International survey panel1, a non-probabilistic census-representative survey panel. For each survey, we sampled at least 730 respondents (15 responses per ad) whose demographics were representative of the U.S. within 5% and who had a range of political views (40% liberal, 40% conservative, and 20% moderate or neutral); across the three surveys we obtained a total sample of 2,886 respondents.

We measured overall ideological divisiveness on the three axes (reporting, approval, and false claims) using two metrics: Within-group divisiveness. Within-group divisiveness measures the extent to which respondents’ answers about a particular ad are consistent with their political ideology. That is, do all liberals answer similarly about a particular ad. For each ad, we first calculate the standard deviation of all the responses, and then we calculate the standard deviation of the responses within a particular ideological group. Next, we compute within-group divisiveness as the fraction of within-group standard deviation to the overall standard deviation. Therefore we interpret values lower than 1 as lower divisiveness (and greater agreeableness) within a group than overall, and values greater than 1 as greater within-group divisiveness than overall.

Between-group divisiveness. Between-group divisiveness measures the extent to which answers from respondents of one political ideology differ from answers of respondents who align with another political ideology. That is, do liberals answer differently about a particular ad than conservatives. For an ad, we calculate the difference between the mean responses per ideological group, and then compute the fraction of this difference over the maximum possible difference given the range of values to obtain the between-group divisiveness measure. This limits the range of between-group divisiveness measure between 0 and 1, where higher values indicate greater divisiveness between ideological groups.

Table 2 summarizes the divisiveness of the high impact ads. We find that the within-group divisiveness measure is lower than 1 for all our surveys. This indicates high agreeableness within the ideological groups. In addition, about 20% of the ads show between-group divisiveness higher than 0.5, indicating severe divisiveness between ideological groups for those ads.

---

1https://www.surveysampling.com/audiences/consumer-online/
(b) Reasons of Inappropriateness

It's violent or prohibited content
It's offensive
It's misleading or a scam
I disagree with it
It's a false news story
It's a spam
Something else

Figure 5: Distribution of the high impact ads on the (a) proportion of reported ads in our dataset, (b) reasons of inappropriateness.

Table 2: Divisiveness measures of the high impact ads.

<table>
<thead>
<tr>
<th>Measure (Group)</th>
<th>Reporting Mean</th>
<th>Stdev</th>
<th>Approval Mean</th>
<th>Stdev</th>
<th>False Claims Mean</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-group divisiveness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liberals</td>
<td>0.87</td>
<td>0.47</td>
<td>0.92</td>
<td>0.36</td>
<td>0.66</td>
<td>0.69</td>
</tr>
<tr>
<td>Conservatives</td>
<td>0.90</td>
<td>0.43</td>
<td>0.98</td>
<td>0.31</td>
<td>0.86</td>
<td>0.63</td>
</tr>
<tr>
<td>Between-group divisiveness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political</td>
<td>0.24</td>
<td>0.18</td>
<td>0.34</td>
<td>0.24</td>
<td>0.17</td>
<td>0.14</td>
</tr>
</tbody>
</table>

3.1 Likelihood of reporting the ads

The first axis of divisiveness that we explored was reporting. We surveyed respondents regarding: 1) Whether they would report the ad shown, and 2) If they would, why do they find the ad inappropriate? Answer choices given, drawn directly from Facebook’s reporting interface [11], were: sexually inappropriate, violent, offensive, misleading, disagree, false news, spam, and something else.

Figure 5 shows the reporting responses for the high impact IRA ads. For over 73% of these ads, at least 20% of the respondents responded that they would have reported the ads. We observe that the majority of the ads were reported on the grounds of being offensive (25%), violent (15%), and misleading (15%). Additionally, a substantial proportion (9%) of the reported responses belonged to the something else category. In such cases, the respondents entered free-text to explain their reason for inappropriateness. Out of the 61 responses that we received in the free-text box, the pre-dominant reasons were that the ad incites racism (20%), and that the ad creates divide (5%) in the society.

Next, to examine ideological divisiveness, we find that the mean within-group divisiveness is 0.87 (stdev = 0.47) for liberals and 0.90 (stdev = 0.43) for conservatives. Both of these within-group divisiveness measures being less than 1, suggests that the likelihood with which individuals within the same ideological group agree about reporting an ad is higher than that when compared against individuals across ideological groups.

3.2 Approving content of the ads

As another characterization of people’s reactions to the ads, we asked respondents in a second survey whether they approve or disapprove of a particular ad, and how strongly they approve or disapprove.11 These questions in the survey were constructed based on questions about political preference that have been extensively pre-tested by Pew Research for previous surveys about political polarization [6]. We find that 87% of the ads were approved and 63% of the ads were disapproved by at least 20% respondents (see Figure 7 (a)). To quantify the received responses, we assigned an approval score on a 5 point scale with values of -2 (strong disapproval), -1 (weak disapproval), 0 (neither approve or disapprove), +1 (weak approval), and +2 (strong approval). While computing the mean approval score for a group, we dropped the 0 responses to ensure

11Specifically, we asked “Do you approve or disapprove of what the ad says or implies?” Answer choices: Approve; Disapprove. Neither; There is nothing in this ad to approve or disapprove of. I don’t know. Followed by a measure of strength “Do you [approve/disapprove] very strongly, or not so strongly?” if the prior question was answered with approve or disapprove.
that a mean approval score close to 0 corresponds to similar weights from approval and disapproval. Table 3 lists some example ads along with their approval tendencies by the two ideological groups within our dataset.

Figures 7 (c&d) show the relationship between respondents’ ideology and approval of ad content. We observe that the mean within-group divisiveness for liberals is 0.92 (stdev = 0.36) and 0.98 (stdev = 0.31) for conservatives (Table 2). Both the within-group divisiveness values being lower than 1, suggests that the likelihood with which individuals within the same ideological group would agree about approving an ad is higher than that when compared against individuals across ideological groups. The divisiveness in approval responses is further confirmed by the between-group divisiveness measure which ranges between 0 and 1 (mean = 0.34) across the high impact ads.

3.3 Perceptions of false claims in the ads
To examine whether the high impact IRA ads contained any false claims, in another survey we asked the respondents if they could identify any false claims present in the ads.\textsuperscript{12} We find that 89% (433 out of 485) of the high impact ads were identified to have at least one false claim, and about 45% of the ads contained false claims according to 10% of the respondents. Figure 8 (a) shows the cumulative distribution of the ads with the number of respondents who identified at least one false claim in them.

Next, as in the other two content analyses, we examined whether respondents’ ideology related to their perception of the presence of false claims (Figure 8). Both the within-group divisiveness values being lower than 1, suggests that the likelihood with which individuals within the same ideological group would agree about finding false claim in an ad is higher than that when compared against

\textsuperscript{12}Specifically, we asked respondents to “Please copy and paste any phrases or sentences in the advertisement that you think contain a ‘false claim’. That is, something that someone could verify as True or False. If you cannot identify any claims, please type ‘No Claims’ in the box.” We then asked them to label the phrases they had identified as ‘True’, ‘False’ or ‘not sure whether they are True or False’.

Table 3: Example ads on the basis of reporting behavior by the respondents from two political ideologies.

<table>
<thead>
<tr>
<th>Group</th>
<th>Approval</th>
<th>Disapproval</th>
<th>Neither</th>
<th>Proportion of Ads</th>
<th>Proportion of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberals</td>
<td>0.92</td>
<td>0.08</td>
<td>0.00</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Conservatives</td>
<td>0.98</td>
<td>0.02</td>
<td>0.00</td>
<td>0.98</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Figure 7: Distribution of the ads on approval and disapproval: (a&b) overall, (c&d) across ideological groups. (a&c) plot the cumulative distribution functions (cdfs), (b&d) plot the differences in approval in each ad, where x-axis consists of all the ads.

Figure 8: Distribution of the ads on false claims (FCs): (a) overall (as a cumulative density function), (b) across ideological groups (where each ad is plotted on the x-axis).
3.4 Summary

This section focuses on peoples’ perceptions of the content of the 485 IRA ads we identified as high impact. To assess these perceptions along three axes – likelihood of being reported, approval and disapproval, and the presence of false claims – we conducted three U.S. census-representative surveys. Our analysis of the perceptions queried in these surveys shows that ideological opinions of individuals influence their perceptions of these ads. We find that many of these ads were relatively divisive, and generated strongly varied opinions across the two ideological groups of liberals and conservatives (see Figure 6, 7, 8).

4 ANALYZING THE TARGETING FORMULA

Next, we focus on understanding how the target formula is created by advertisers and the role that Facebook interface plays on that.

4.1 Targeting Possibilities

The Facebook ads platform provides three approaches for advertisers to target people [3, 23], briefly described next.

Personally Identifiable Information (PII) targeting is the form in which advertisers provide personal information about users such as name, phone number, and email address so that Facebook can directly place the ads to them. This kind of targeting does not appear in the IRA dataset.

Look-alike audience target. For this targeting option, advertisers provide to Facebook a list of users similar to that one in the PII or a list of people who liked the advertiser Facebook page. Then, Facebook attempt to target a similar audience to the group in this specific list. Only 1.1% of the high impact ads used this option.

Attribute-based targeting allows the advertiser to create a target formula based on a wide range of elements that include user basic demographics (i.e. gender, age, location, language), advanced demographics (i.e. political leaning, income level, "Parents with children preschoolers"), interests (i.e. newspapers, religion, politics), and behaviors (i.e. ‘Business Travelers’ or ‘New Vehicle buyers’). Recent work showed that the number of possible interests provide by Facebook is greater than 240,000 [23]. Facebook allows one to include or exclude users with each of those attributes and combine multiple attributes as part of a target formula. The vast majority of the high impact ads, 895 out of the 905, used this option to elaborate a formula. We found that 78% of the ads used 2 or more interests and behaviors in their formula, creating very complex formulas with up to 39 distinct attributes.

Figure 9 shows the top attributes that appear in the ads target formula based on the number of times they appeared in different ads. There were 497 distinct attributes and the most present attributes interest were African-American history and African-American Civil Rights Movement (1954-68), appearing in 295 (32%) ads. We can note a prevalence of attributes related to African-American and Hispanic Population, with interests like Mexico, ‘Hispanidad’ and ‘Latin hip hop’. Next, we investigate aspects of the Facebook ads platform design that might have favored the IRA ads to massively explore this particular targeting strategy.
4.2 The Role of Attribute Suggestions

Facebook provides a tool for advertisers that, given a target attribute, it presents a list of other attributes that target people with similar demographic aspects [23]. For example, in the list of suggested targeting interests for ‘Townhall.com’, a page with an audience in which 79.5% of the users are very conservative users according to Facebook, there are other pages with similar bias towards very conservative users, i.e. ‘The Daily Caller’ (67.1%), ‘RedState’ (84.3%), and ‘TheBlaze’ (59.6%) [20].

In order to investigate if the IRA ads have used suggestions to elaborate complex targeting formulas, we crawled the attribute suggestions for each attribute that appear in the dataset of highly impact ads. Figure 10 shows the cumulative distribution function for the number of suggested attributes that appear in the same formula. We can see that around 64% of the ads that potentially used this feature because they have at least three target attributes suggested by Facebook as part of the same formula. There are 1.2% of ads with more than 10 suggested attributes in the same formula. As an example, all the 13 interests, including Islam, Ramadan, Islamism, used in the target formula of the ad ID 1915\textsuperscript{13} appear as suggestions for at least one of the others in the formula. For ad ID 1840\textsuperscript{14}, we were able to find 9 out of 10 of the interests using the interest suggestion feature. This provides evidence that this feature may have been a key element used by the IRA campaign to choose the target audience.

4.3 Summary

In this Section we show that the vast majority of the IRA ads use attribute-based targeting, containing complex target formula that includes interest and behavioral attributes that are likely suggested by Facebook. Next, we investigate the extent to which these formulas allowed advertisers to reach demographic biased audiences.

5 ANALYZING THE TARGET AUDIENCE

We start by describing our methodology to reproduce the IRA queries (without running the ad) and gather the demographics of the of the targeted users.

5.1 Assessing the Audience Demographics

Before launching an advertisement in Facebook, the advertiser can get the estimated audience (i.e., the number of monthly active users) likely to match the target formula. Our methodology consists of using the Facebook Marketing API\textsuperscript{15} to reproduce the targeting formula of all high impact IRA ads and get the demographics of the population that matches each targeting formula, without running any ad. This methodology has been extensively used recently for different purposes, including inferring news outlets political leaning [20], study migration [26] and gender bias [13] across countries, and for public health awareness [21] and lifestyle disease surveillance [5]. For our analysis, we considered seven demographic categories: political leaning, race, gender, education level, income, location (in terms of states), and age. As a baseline for comparison, we also gathered the demographic distribution of the United States Facebook population.

Only 11% of the used attributes that appear in the IRA ads targeting formulas are not available for targeting anymore due to changes in the Facebook Marketing API. In most of these cases, we reproduced the ad target formula without the missing attribute, especially when the attribute looks redundant with the others in the formula. We did not reproduce only 6 targeting formulas.

5.2 Measuring Audience Bias

To assess the audience bias of each of the demographic aspects that we considered, we computed the differences between the fraction of the population with a demographic aspect and the same fraction of the population in the baseline distribution (i.e. the U.S. Facebook population), namely the bias score. For instance, if the percentage of African-Americans in the audience of a particular ad is 40%, the bias score for this dimension in the ad is 0.25 as the percentage of African-American in the U.S. Facebook population is nearly 15.5% (0.4 – 0.155).

Figure 11 depicts the distribution of the measured bias on political leaning and ethnic affinity. In comparison with all the demographic categories, these two showed to be the ones with the highest biases.

\textsuperscript{13}http://www.socially-divisive-ads.dcc.ufmg.br/app.php?query=1915

\textsuperscript{14}http://www.socially-divisive-ads.dcc.ufmg.br/app.php?query=1840

\textsuperscript{15}developers.facebook.com/docs/marketing-apis
We note that most of the ads target audiences that are more biased towards the African-Americans population and the Liberals. More specifically, about 70% of the IRA ads target an audience with a higher proportion of African-Americans than in the US Facebook distribution. This difference is even accentuated for Liberals, with 82% more biased in comparison with the reference distribution. The percentage of ads with bias score superior to 0.15 is 52% for African-Americans specifically, about 70% of the IRA ads target an audience with a higher proportion of African-Americans than in the US Facebook distribution. This difference is even accentuated for Liberals, with 82% more biased in comparison with the reference distribution. The percentage of ads with bias score superior to 0.15 is 52% for African-Americans than in the US Facebook distribution. This difference is even accentuated for Liberals, with 82% more biased in comparison with the reference distribution. The percentage of ads with bias score superior to 0.15 is 52% for African-Americans.

5.3 Targeting audience and Divisiveness

Next, we investigate if the advertisers target the ads towards audiences that are less likely to identify their inappropriateness due to their ideological perception bias. Additionally, we examine if the ads directed to biased audiences could leverage the already existing societal divisiveness to further amplify it among the masses.

To understand these nuances of targeted advertising, in this section, we focus on the relationship between the targeted population and the ideological divisiveness in reporting, approval, and false claim identifying behaviors for the ads. Table 6 reports the correlation values between the targeted population and the tendency of the population to report, approve, and identify false claims.

We observe a negative correlation in the case of reporting for both Liberals and Conservatives (also see Figure 12 (a)).

Table 6: Pearson’s r correlation between targeting and the ideological divisiveness for the high impact ads (*** p < 0.001, no statistical significance in the case of false claims).

<table>
<thead>
<tr>
<th>Group</th>
<th>Report</th>
<th>Approval</th>
<th>False Claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberals</td>
<td>-0.17***</td>
<td>0.41***</td>
<td>-</td>
</tr>
<tr>
<td>Conservatives</td>
<td>-0.15***</td>
<td>0.32***</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 11: Bias in demographic dimensions. Each violin represents the bias score for all high impact ads in a particular demographic dimension. The median is represented by a white dot in the center line of the violin. 50% of the data is present between the two thick lines around the center.

Figure 12: Relationship between targeting and the responses by ideological groups. (a,c,e) show the proportion of population targeted and their tendency of response. Each circle represents an ad, and their size is proportionate with the between group disputability for that ad. (b,d,f) compares the mean responses of the targeted ads with their hypothetical non-targeted counterpart (i.e., overall responses), where each ad is represented on the x-axis.

This suggests that the targeted population has a lower tendency to report than the non-targeted one. This is also evident per Figure 12 (b), where we find that the reporting by the targeted population carries way lower likelihood than the reporting by the overall (or non-targeted) population.

Approval. We observe a positive correlation in the case of approval for both Liberals and Conservatives (also see Figure 12 (c)). This suggests that the targeted population has a greater tendency to approve the ads as compared to the non-targeted population.
is also evident per Figure 12 (d), where we find that the approval score by the targeted population carries greater score for a majority of the ads compared to the overall (or non-targeted) population.

False claims. For false claims, we do not find any significant correlation between the targeted population and divisiveness. However, per Figure 12 (e&f) we do find that the targeted population has a lower tendency to identify false claims.

Taken together, we can assume that the ads were “well-targeted” in a way towards that population which was more likely to believe, and approve and subsequently less likely to report or identify false claims in them.

5.4 Summary
Our findings show that the IRA ads reached audiences that are very biased towards African-Americans and Liberals. More important, we show that ads were overall targeted towards a population that is more likely to believe, and approve and subsequently less likely to report or identify false claims in them.

6 CONCLUDING DISCUSSION
In this paper, we provide an in-depth quantitative and qualitative characterization of the Russia-linked ad campaigns on Facebook. Our findings suggest that the Facebook ads platform can be abused by a new form of attack, that is the use of targeted advertising to create social discord. These ads showed to be divisive, were 10 times more effective than a typical Facebook ad, were biased especially in terms of race and political leaning, and tended to be targeting more the users who are less likely to identify their inappropriateness. We also provide strong evidence that these advertisers have explored the Facebook suggestions tool to engineer the targeted populations.

While this tool may be helpful in many ways, it needs to be carefully redesigned to avoid that a malicious advertiser reaches so easily groups of vulnerable people. For example, Facebook recently presented its intention to manually inspect ads before they are launched [1], aiming to guarantee that ads do not divide or discriminate people. Our work suggests that the priority of the candidates to be manually inspected can be based on their targeting formula. For instance, those ads that target extremely narrowed populations, on the basis of race, political leaning, and other sensitive topics have greater likelihood of being divisive. Additionally, the ads that experience severely high click-through rates could also be flagged to be quickly inspected.

As a final contribution, we have deployed a system (available at http://www.socially-divisive-ads.dcc.ufmg.br/) that displays the ads and their computed information such as the demographics of their targeting audiences.

7 ACKNOWLEDGMENTS
F. Benevenuto and F. Ribeiro acknowledge grants from Capes, CNPq, and Fapemig. E. M. Redmiles acknowledges support from the U.S. National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1322106 and from a Facebook Fellowship. This research was partly supported by an European Research Council (ERC) Advanced Grant for the project “Foundations for Fair Social Computing”, funded under the European Union’s Horizon 2020 Framework Programme (grant agreement no. 789373). This research was partly supported by ANR through the grant ANR-17-CE23-0014.