Inauthentic Newsfeeds and Agenda Setting in a Coordinated Inauthentic Information Operation

Carl Ehrett, Darren Linvill, Hudson Smith, Patrick Warren, Leya Bellamy, Marianna Moawad, Olivia Moran, and Monica Moody

Clemson University

Abstract

The 2015-2017 Russian Internet Research Agency's coordinated information operation is one of the earliest and most studied of the social-media age. A set of 38 city-specific inauthentic "Newsfeeds" made up a large, under-analyzed part of its English-language output. We label 1000 tweets from the IRA Newsfeeds and a matched set of real news sources from those same cities with up to five labels indicating the tweet represents a world in unrest and, if so, of what sort. We train a natural-language classifier to extend these labels to 268k IRA tweets and 1.13m control tweets. Compared to the controls, tweets from the IRA were 34 percent more likely to represent unrest, especially crime and identity danger, and this difference jumped to about twice as likely in the months immediately before the election. Agenda-setting by media is well known and well-studied, but this weaponization by a coordinated information operation is novel.

Keywords*: Agenda Setting, Disinformation, Internet Research Agency, Social Media, Twitter*

The Russian backed Internet Research Agency's (IRA) coordinated disinformation operation worked to influence the 2016 U.S. Presidential election, in large part through social media (Jamieson, 2018). While fairly well studies, a major piece of this operation has remained relatively unexplored and its purpose left unknown. Here we will demonstrate how the IRA strategically employed real local news for agenda setting purposes in an apparent effort to show Twitter users a world more dangerous and unrestful than they may otherwise experience.

Complex state-backed coordinated information operations often consist of several distinct sets of accounts, each playing a specialized role. In the context of the IRA, Linvill and Warren (2020b) identified five primary account types: Left Trolls, Right Trolls, Newsfeeds, Fearmongers, and Hashtag Gamers. Of these, the Left Trolls and Right Trolls have received the most attention (Bastos & Farkes, 2019; Freelon et al, 2020). But from the perspective of overall output, Newsfeeds made up a substantial part of the operation, actually originating more English-language tweets than any other type and producing 26% of the English-language tweets in 2016. Nonetheless, there exists no detailed analysis of the role these accounts played in the operation.

The bulk of Newsfeed content emanated from 38 accounts that operated from early 2015 to mid-2017, each of which purported to be a local-news provider for one of 36 large cities around the United States.[1] Figure 1 illustrates how two of these accounts appeared when active. In general, these accounts gathered and reposted news from a small set of legitimate local news sources that served the targeted city, mostly local newspapers and television. Newsfeeds were largely automated, using clients like Twitterfeed and Twibble to gather stories from RSS feeds or Twitter feeds to repost as their own content. They made very limited use of retweets, but often linked to

[1] 16 additional accounts classified as Newsfeeds were specialized, topic oriented accounts such @TodayInSyria. They are not analyzed here.

the legitimate website from which the news content originated. They did not post grossly inaccurate information (like Fearmongers), link/amplify the more ideological troll accounts or extremists outside the network (like Right or Left Trolls), or seem to attempt to contribute to the trending of topical hashtags (like Hashtag Gamers).

Not only was Newsfeeds' behavior inconsistent with other IRA types, it also lacks comparison with the tactics identified in any other major coordinated information operations. They did not flood a conversation to make real voices hard to find, as has been common in many campaigns (Roberts, 2018). Instead, they posted the real news of the day. They did not focus messaging on key dates (King, Pan & Roberts, 2017; Keller, et al, 2020), instead they posted quite regularly, almost mechanically. They did not connect with interest groups to drag them to more extreme positions; instead they hardly interacted with other accounts at all.

As we demonstrate, the Newsfeeds seem to have attempted a form of agenda setting by disproportionately highlighting some elements of the real news. This biasing could serve a variety of purposes, including making emphasized subjects more prominent in the minds of the Newsfeeds' followers. After all, authentic news media has been shown to have such an agenda-setting impact (McCombs & Shaw, 1972; Iyengar, Kinder, & Peters, 1982; Dellavigna & Kaplan, 2007). We hypothesize that the Newsfeeds' mission was to present a biased reflection of American life, one that was more rife with unrest, conflict, and division than traditional news would present.

Our approach is to measure the share of tweets that presents a world beset with unrest, and if so, what sort of unrest (from 5 sub-categories). We then contrast these rates between the Newsfeed accounts and those in a matched sample of actual local media operating in the same city from which the Newsfeeds were drawing the majority of their content. For a sample of tweets from each group, we code by hand whether each tweet portrays unrest. We use this hand-

coded dataset to train a machine-learning classifier, which we validate and use to extend our classification, obtaining confidence scores for each tweet in the full dataset of 1.4 million tweets. With these scores applied to every tweet, we measure the differences in the picture of the world that the IRA presents to the picture of the world the legitimate local media sources present on Twitter, across cities and over time.

We find large and significant differences in the level of unrest that these two sets of accounts portray. These differences increase dramatically in August 2016. Prior to that time, we find that IRA Newsfeeds have 10% higher odds of containing unrestful content than matched accounts controlling for location by month fixed effects, whereas for the period beginning in August 2016 they have 100% higher odds of unrestful content. Over the full time period studied, IRA Newsfeed had about 34% higher odds of sharing unrestful tweets than their matched control accounts, and that difference is driven by unrest related to crime, institutional failure, and identity danger, rather than acts of god. This pattern is consistent with an agenda-setting strategy with a goal to make the United States seem a riskier and more violent place than it really is, in the months leading up to the 2016 U.S. Presidential Election.

Agenda-Setting and the IRA

The agenda-setting theory suggests that while news media may not tell audiences what to think, they can tell audiences what to think about (McCombs & Shaw, 1972). The two basic assumptions of what is called first-level agenda setting are, first, the media shape reality rather than simply reflecting it and, second, the more attention media focus on an issue the more likely the public is to feel it is important to society. Although these assumptions suggest powerful effects,

agenda setting is a theory of limited media effects with individual issue relevance being a fundamental contributing factor (Erbring, Goldenberg, & Miller, 1980; McCombs, 1994).

Second-level agenda setting links the theory with framing and suggests media attention can influence how people think about a topic by ignoring or downplaying certain attributes while selecting and emphasizing others (Ghanem, 1997). In effect, media tell audiences *how* to think about issues. Kiousis et al. (2006) explain second-level agenda setting through an example addressing candidate images, saying, if media "emphasize the integrity of a political candidate in news stories, public descriptions of that candidate should also stress his or her integrity" (p. 269). Local news in particular, being both popular and trusted, has been found to have just such an agenda setting effect on issues related to crime (Gross & Aday, 2003).

The internet, and social media in particular, has dramatically changed the media landscape since agenda-setting was first conceptualized. McCombs (2005) discussed the argument that the internet will lead to the end of agenda setting as individuals have more personally tailored media available to them and gain the ability to choose from a wider array of online news and information. It is possible, however, that far from limiting the ability of media to set an agenda, because of its ability to reach targeted audiences social media can be a tool expressly for agenda-setting purposes.

With this in mind, agenda-setting theory has been used by scholars as a lens to better understand the potential of misinformation and disinformation. For example, Dreier and Martin (2010), employed agenda-setting as a lens to examine controversy surrounding the community group ACORN and found a persistent, online campaign with limited truth can influence the national news media. Vargo, Guo, and Amazeen (2018) utilized intermedia agenda-setting theory (examining the interaction between media outlets in setting each other's news agenda) to better understand the agenda-setting power of fake news, finding it was particularly influential among

partisan media. Finally, Pierri, Artoni, and Ceri (2020) used agenda-setting to better understand disinformation spreading on Twitter proceeding the 2019 European Parliament elections. They found that a small number of websites, while having a limited impact on broader online discussions, had an outsized impact on far-right political discourse on the platform.

The documented behavior of the better understood Right and Left Trolls makes clear that an overarching goals included instigation of conflict and unrest (DiResta, et al., 2019, Jamieson, 2018; Linvill et al., 2019; Linvill & Warren, 2020b). Since agenda setting is an important mechanism by which news media may affect citizens' perception of conflict (McCombs & Shaw, 1972; Iyengar, Kinder, & Peters, 1982; Dellavigna & Kaplan, 2007), it is plausible that the IRA attempted to use their Newsfeed accounts for this purpose, which leads to our first hypothesis.

H1. *Compared to their matched control sources, the IRA Newsfeeds include more news that represents a society at unrest.*

More detailed investigation of the themes highlighted in the Right and Left Troll accounts' output indicates that they were particularly likely to include attacks that were based on identity (Arif et al., 2018; Freelon et al., 2020), investments to build capital within identity groups (Linvill & Warren, 2020a), or which tried to undermine trust in institutions (Linvill, et al., 2019). Our second hypothesis says that we expect similar strategies from these accounts.

H2. *Compared to their matched control sources, the IRA Newsfeeds include more news that highlights unrest related to human conflict, such as identity danger, crime, or the failure of institutions, rather than conflict/unrest due to acts of nature.*

Finally, the IRA's strategy implemented on their Left and Right Troll accounts has been shown to have varied dramatically over time, with sudden shifts in behavior as accounts moved from a "Growth" phase to an "Amplification" phase in the month before the 2016 Presidential

election (Linvill & Warren, 2020a). A parallel shift in the behavior of the Newsfeeds is plausible, leading to our third hypothesis.

H3. *The agenda-setting behavior of the IRA Newsfeeds shift substantially at some point immediately prior to the 2016 election, becoming more extreme as the accounts move from a growth to an activation phase.*

Methods

The raw data for this analysis come from two sources. First, we collected the output of the 38 locally oriented IRA Newsfeed accounts from Twitter's January 2019 update to their October 2018 release of the output they linked to the IRA (Roth, 2019), for the 13 months running from December, 2015 to December, 2016. We supplement these with 60,975 additional tweets produced from these same accounts during this period and downloaded from Social Studio but not included in the Twitter release.[2] These accounts are the subset of the accounts identified as Newsfeeds by Linvill & Warren (2020b) that explicitly identify as local news providers. This results in 267,745 tweets, split among 38 accounts, from very small @ChesterCityNews (269 tweets) to a very large @KansasDailyNews (22,197 tweets). As not every local Newsfeed was active every month, this results in 407 Newsfeed-month groupings, instead of the 494 (38 x 13) combinations one would observe if all accounts were active in all months.

For each of these Newsfeeds, in each month, we identified the general Twitter news feeds of the real local sources from which the Newsfeed pulled their news, to provide a "control group" for each Newsfeed. In many cases, this identification simply came from tabulating the domain of the links provided in the IRA tweets. But it sometimes required stepping through a link shortener/client, such as Twibble.io (in 57 out of 407 Newsfeed-months) or searching Twitter or

[2] It is not clear why these Tweets were not included in the original Twitter release, but the most likely explanation is that they were deleted by the accounts before the accounts were shut down by Twitter.

Google for the text contained in a sample of their tweets in order to identify the original news provider (in 188 out of 407 cases). When multiple sources were identified, the 3 most common sources for the month were chosen. Multiple sources were identified in 157 of the 407 cases. Once these original sources were identified, the full text of their most general news Twitter account was collected for that month. This resulted in 1.13 million "control" tweets from 95 accounts.[3]

These accounts are meant to represent how a real local news source would have reported about the events in the city. To the extent that they bias coverage, relative to reality, we are estimating the additional bias injected by the IRA. By choosing to use the specific set of accounts from which the IRA pulled their news (rather than randomly selected local news accounts), we are avoiding the possibility that any bias we observe is driven by the IRA selecting relatively extreme sources. But casual inspection of their sources makes that explanation implausible, as they consist primarily of major local newspapers and network broadcast affiliates.

Qualitative Analysis

From the full data set of ~1.4 million tweets, both from troll and legitimate news sources, we drew a stratified random sample of three troll account and three control account tweets for each unique city-month pair in order to account for systematic variation across time, locale, and account type (troll or control). This resulted in a set of approximately 1800 tweets. Of these, 1000 tweets were randomly sampled for hand coding. These tweets were labelled by hand with five binary, non-mutually exclusive categories: "identity danger", "institutional failure", "act of god", "crime", and "other unrest" as defined below. Any tweet that received any of these labels inherited the umbrella label of "any unrest" and all others were coded as "not unrest".

[3] Due a programming error, the controls for @TODAYPITTSBURGH were not collected until July, the controls for @OAKLANDONLINE were not collected in Jan-Mar, and the controls for @CHESTERCITYNEWS, @CAMDENCITYNEWS were not collected in December, 2015. These city-months are dropped from the analysis.

We conducted qualitative analysis as recommended by Corbin and Strauss (2015). First, we read posts to get a sense of the data and then engaged in unrestricted coding, working together to compare and conceptualize data. From this process we identified meaningful patterns. Second, we conducted axial coding by identifying linkages between ideas underpinning individual tweets. This process reduced patterns into categories. As we continued, we identified sample tweets for each category and created definitions for each category to help clarify meaning.

To maximize the reliability of our analysis we created and employed a code book. The use of a code book served as a stable representation of the analysis and later served as a reference throughout the coding process (Creswell & Poth, 2018). Members of the research team coded common, randomly selected sets of tweets. After each set was coded we compared results and refined our analysis. This process was continued until we met an acceptable Krippendorff's alpha reliability of 0.70 (Krippendorff, 2004). The random sample of 1000 tweets was then analyzed. Tweets were coded for each category as either present or not present, an unrest label applied to any tweet that had a high likelihood of arousing a state of dissatisfaction, disturbance, or agitation in a group of people. Tweets were further labelled using one or more of the subtypes, below. Multiple categories could be present simultaneously.

Identity harm. These tweets overtly mention danger for a specific identity group. Identity was defined broadly and groups included religion, race, ethnicity, political group, gender, sexual orientation, nationality, age, socioeconomic class, or veteran status. Danger was considered the threat of immediate or future risks or harm relayed. Posts for which it was decided an informed reader could connect the danger mentioned to the identity group mentioned were placed in this category. Tweets placed in this category included the August 17, 2016 tweet “Parents ramp up

attack on transgender rules in Fort Worth schools" and the October 9, 2016 tweet "Should you watch the second presidential debate w. your children? v. @stltoday #Debates2016".

Institutional failure. Tweets in this category mention the failure, decline, or collapse of formal or informal institutions. Institutions included a law, practice, custom, or organization that is important to the functioning of mainstream society in a country or community. It included, but was not limited to, institutions such as courts, law enforcement, elections, media, specific industries, family, and organized religion. Example tweets in this category included the February 6, 2016 tweet "Bridgestone-Firestone recalls over 36,000 truck tires" and the December 20, 2015 tweet "Here's How the #FederalReserve Just Told #Black Folks 'You Don't Really Matter'".

Acts of God. Tweets in this category included accidents or events not directly caused by humans. It included posts that addressed weather and other natural occurrences but also accidents, such as automobile accidents, which were not described as the direct result of human negligence. This included tweets such as the December 12, 2016 tweet "A storm is approaching SoCal - track it with the LIVE Megadoppler 7000 HD" and the July 1, 2016 tweet "Your comments: Be wary of pit bulls, one reader writes".

Crime. This category included tweets that specifically mentioned a crime. In addition to notices of recent crimes occurring, it also included discussion of arrest, conviction, or sentencing and other court related matters. It included posts such as the January 1, 2016 tweet "Cosby files motion to dismiss sex assault charges" and the September 9, 2016 tweet "Couple charged with killing woman and kidnapping her children".

Other. This category was reserved for tweets for which unrest was clearly addressed but which did not fit into one of the other five categories; this was generally because not enough

context was given in the text of the tweet. This included tweets such as the March 24, 2016 tweet "UPDATE: The lockdown at Naval Medical Center San Diego has been lifted."

The remaining tweets received no flags. These tweets came in a range of forms. For example, many of these posts involved news about sports and entertainment, such as the January 31, 2016 tweet "AFL exhibition games in #Fort Worth helped shape pro football landscape". Others were news of events surrounding the election, including the October 3, 2016 tweet "Clinton Visits Charlotte Church, Calls for Healing #politics".

Training the Classifiers

Using the labeled set of hand-coded tweets we trained logistic regression models for each of the six binary labels (overall unrest, plus the 5 subtypes). These models use a vector of features about the content of the tweet to estimate a predicted probability that a human coder would assign the specified label (Pampel, 2000). Prior to modeling, the tweets were pre-processed to convert all characters to lowercase, expand contractions, remove non-standard utf8 characters (such as emoji), tokenize hashtags and miscellaneous symbols, and remove slashes, brackets, punctuation and stopwords. The remaining text is then converted to a numerical representation, which forms the core of the independent variables in the logistic regression.

There are several possible methods for converting textual data into numerical inputs for use in the logistic regression. The most familiar of these is bag-of-words (BoW) count vectorization (and related variants). BoW has no sense of the relationships between words. For example, in logistic regression, the effects of the words "GOP" and "Republicans" would be modeled with independent parameters. This makes the count vectorization method inefficient when dealing with small datasets and short texts. For this reason, we make use of a more contemporary embedding technique commonly known as Word2Vec (Mikolov, 2013). In

Word2Vec, the unique words in a corpus are each mapped onto a k-dimensional vector (we choose k=100) of numbers such that words appearing in similar contexts have similar vectors. Since this mapping criteria makes no reference to the class labels, we are able to train this embedding model on the full dataset, thereby incorporating more of the particular linguistic conventions in our corpus than if we were restricted to our labeled set. With the word embedding model in hand, the numerical representation of a tweet is calculated as the average of the normed vectors for each word in the tweet. The numeric features produced by this process theoretically form a more efficient representation for classifying Tweets as compared to more contextually blind approaches. We used the implementation of Word2Vec in the open source fastText library which includes several technical improvements over the original (Bojanowski et al. 2017, Joulin et al. 2016).

Using a training set of 85% of the labeled data stratified with respect to "unrest" (so that approximately the same proportion of unrestful tweets appear in both the training and test sets), we used grid search cross validation to tune the hyperparameters of the logistic regression models. We then evaluated the classifiers' performance using the hold-out set of 15% of the labelled data. The resulting micro-averaged ROC score is 0.92, indicating good overall performance. The relevant ROC curves appear in Figure 2, along with the area under the ROC curve for each class. Figure 3 shows the precision/recall curves along with the maximum F1 score achieved by each classifier on the holdout set. Table 1 shows some examples of the model output for particular tweets, both in cases in which the model was successful and in the (infrequent) cases of model failure. One can see that all classifiers achieve good performance with the exception of "other unrest". This is likely due to the low prevalence of positive cases of this category -- only 12 in the set of 1000 labelled tweets. We therefore exclude this category from further analysis. Prior to undertaking the regression analysis, we retrain the classifiers on the full labelled data set.

Regression Analysis

For each tweet in the full data set, we generated six labels using the six classifiers trained on the hand-labeled data set. Each label is a probability score; e.g., for a given tweet the label for "unrest" is a numeric value in the range [0,1] which is the probability that the tweet contains unrestful content. We then apply the logit transformation to induce the labels to more closely approximate a normal distribution. The resulting dependent variables are thus the log odds, for each category, that a human coder would assign the relevant label.

The independent variables for the analysis include one binary variable, "is IRA", for each tweet indicating whether or not the tweet originates from an IRA account. We also label the data with 407 dummy variables corresponding to the 407 unique combinations of month and city represented by the tweets in the data set. For IRA Newsfeeds, the city of a tweet is determined by the city that the Newsfeed was purported to represent. For the control accounts, the city was defined as the city that the associated IRA newsfeed account was purported to represent. The city-month dummies are included in the regression to control for the effect of systematic bias due to time or locale. E.g., if IRA tweets tended to be concentrated in locales that were more unrestful -- such as high-crime cities -- then IRA tweets might be on average more unrestful than legitimate news feed tweets simply due to this locale bias. Including the dummy variables controls for this, so that we can learn whether IRA tweets are more unrestful than legitimate feed tweets from the same time and locale.

Formally, for tweet i, in city c, in month t, we estimated variants of regressions of the form

$$Unrest^{d}{}_{ict} = \beta^{d} isIRA_{ict} + \delta^{d}{}_{ct} + \epsilon^{d}{}_{ict},$$

where $d \in \{Any, Identity, Crime, Act\ of\ God, Instit.Fail, Other\}$ represent the different definitions of unrest, $Unrest^d$ represents the log odds that that the tweet would be coded as containing unrest of type d, ${\delta^d}_{ct}$represent the month-city dummies, and ${\epsilon^d}_{ict}$is an error term. To estimate the standard error on our estimate of this coefficient, we clustered at the city-month level. To accommodate the size of the data set, we relied on the biglm (Lumley, 2020) and bigcluster_sandwich (Tsay, 2013) R packages to fit the model and find the cluster-robust coefficient covariance (MacKinnon & White, 1985) respectively.

Our primary interest is in β^d, the coefficient on the isIRA dummy, which is understood in connection with the fact that the dependent variable is the log odds of category membership. The regression coefficient is the log of the ratio of the odds of category membership for IRA tweets to the odds of category membership for non-IRA tweets. E.g., in the case of "crime", in a given time/locale, the estimated regression coefficient of 0.17 for "is IRA" indicates that that the odds of an IRA tweet being labelled as "crime" are estimated to be $exp\ (0.17) \approx 1.19$ times the odds of a non-IRA tweet being so labelled. The reported p-values employ the (very conservative; Moran, 2003) Bonferroni correction for the multiple simultaneous tests performed.

The overall average gaps in this regression potentially elide important differences over space and time in the agenda-setting behavior of the Newsfeed accounts. Differences across cities or over the months of the campaign could exist for two reasons: The aims of the campaign could vary, or the aims could be constant but the most efficient means of achieving those aims could vary. To explore this, we also perform variants of the regressions, in which we allow the coefficient on the "is IRA" dummy to vary over space or time.

Sentiment/emotion Analysis

In addition to the above-described analysis using logistic regression classifiers trained on our own hand-labeled data, we also classify the full data set using IBM Watson's Natural Language Understanding (WNLU) tool. Using this tool, we obtain one "sentiment" score (a value from -1 to 1 which indicates whether the sentiment is positive, neutral, or negative), and five emotion scores (each a value from 0 to 1 indicating how likely the tweet is convey the relevant emotion): anger, disgust, fear, joy, and sadness. Similar to the above analysis of unrest categories, we apply the logit transform to these scores. In this case, prior to the logit, we translate all the scores to the interval $(\epsilon, 1-\epsilon)$ where ϵ is the (extremely small) machine epsilon value of the system; this is due to the fact that scores of 0 or 1 would otherwise be logit-transformed to negative or positive infinity, respectively. Having obtained the logit-transformed sentiment and emotion scores, we apply the same regression analysis described above.

Results

In the first two columns of Table 2, we present the shares of tweets labelled with each of the five substantive unrest labels, for both the IRA Newsfeeds and the controls. In the third and fourth columns, we present the mean predicted probability of each unrest label being assigned by our classifier for the full datasets IRA Newsfeeds and controls. For each category, a z-test statistics of the difference in mean probability between the IRA Newsfeeds and the controls for the full time period studied is reported in the final column, and they are all significant at (well below) the 0.001 level.

Both overall and for each sub-category that represent human-originating problems, the IRA produced significantly more unrestful content than did the matched sample of controls, particularly beginning in August 2016. In that time period, the overall difference in any form of unrest was approximately 14.1 percentage points. The IRA trolls produced approximately 5.4 points higher

share of tweets with identity danger; approximately 2.3 points higher share of tweets with institutional failure; and approximately 6.1 points higher share of tweets with crime. Only for unrest representing acts of god was there no substantive difference between the trolls and controls in the period beginning in August 2016.

Basic Regression Results

The patterns observed in the mean differences are robust to controlling for city-by-month fixed effects, in a regression, and comparing label shares within those clusters. The estimated coefficients on the "is IRA" dummy and associated p-values, in the regressions explaining modeling estimated probability of each label, appear in Table 3. Re-running the regression with city-by-week fixed effects produces nearly identical results. In the case of the unrest labels, we find "is IRA" to be significant at the 0.001 level for each of the five labels analyzed. IRA Newsfeed tweets have higher odds than control news tweets to be labelled unrestful (34% more), and to refer to identity danger (24% more), crime (19% more), and institutional failure (11% more). They are have lower odds of referring to acts of god (16% less).

The results are similar for the WNLU sentiment/emotion analysis. There, we find that "is IRA" is significant for each of the five emotion labels analyzed, at far below the 0.01 level. For sentiment, "is IRA" is significant at the 0.05 level prior to correction for the six-way family-wise error rate, but not with this correction. The relevant regression coefficients and p-values appear in Table 4. IRA tweets have higher odds than legitimate news tweets to contain anger (22% more), disgust (51% more), fear (7% more), and sadness (28% more), and lower odds of joy (12% less).

Heterogeneous Gaps

The mean overall contrasts between the IRA Newsfeeds and the controls are large, but they elide significant heterogeneity across space and time. Figure 4 presents the results for each city

that IRA Newsfeeds purported to serve, where the horizontal position of the dot indicates the estimated difference in unrest shares, on average for indicated city, the size of the dot indicates the number of tweets from IRA Newsfeeds targeting that city, and the whiskers indicate 90% confidence intervals around that estimate. The first panel, for “Any Unrest” is sorted by the size of the estimated difference in that outcome between IRA Newsfeeds and the control outlet(s) in that city. The IRA Newsfeeds present a significantly more unrestful world in 24 of the cities, significantly less unrestful in 6, and no significant difference in 7. The role of geographic targeting, if any, is not entirely clear. All 4 Texas cities are among the set with substantial unrest contrasts, as are all three in Pennsylvania, and both in New Jersey. The largest contrast is in Richmond, VA. In contrast, Wisconsin has both the largest anti-unrest contrast (Minneapolis) and a substantial pro-unrest contrast (Milwaukee).

The other 5 panels present the results for the five specific varieties of unrest, where the order the cities are listed in is held constant (in order of estimated “Any unrest” contrast). The distribution of coefficients for the "Crime" and "Identity Danger" contrasts lines up fairly well with each other and with the "Any Unrest" results, in terms of what cities have large differences, small differences, and negative differences. A similar but less pronounced pattern continues for "Other Unrest", with even less consistent results for "Institutional Failure" and "Acts of God".

To look at variation in the estimated coefficients across time, Figure 5 presents how the results vary by the month in our sample, where the vertical position of the dot indicates the estimated difference in log unrest label odds, averaged across cities, the size of the dot indicates the number of tweets produced by IRA Newsfeeds that month, and the whiskers indicate 90% confidence intervals around that estimate. These graphs show dramatic increases in the relative level of unrest presented by IRA Newsfeeds beginning in August, 2016 and persisting (and, in some cases,

increasing) through the end of the year. For overall unrest, it jumps from 10% higher odds of unrest to 100% higher. This jump occurs across nearly all the unrest types, with the sole exception of Institutional Failure, which increases more smoothly, but the largest increases occur in the Crime and Identity Danger categories. Finally, there is a smaller, but statistically significant, jump of Identity Danger contrast that precedes the bigger jump, beginning in June, 2016. At the same time that the share of unrest jumps, the overall output drops, perhaps indicating that biasing happens through the removal of innocuous content rather than the insertion of restful content. Examining the data with an August starting point, the IRA is significantly (alpha = 0.05) more unrestful in 27 of the 33 cities, significantly less unrestful in two cities (Detroit and Oakland), and have no significant difference in four of the cities (San José, Jackson, New Orleans, and St. Louis).

But these time series patterns are not independent from the geographic ones, as IRA Newsfeeds are not all equally active every month. To summarize these interactions, Figure 6 presents point estimates by city and month for each of our unrest measures. In it, darker, purple, squares represent city-month pairs in which the IRA Newsfeeds produce tweets that we estimate to be more likely to be unrestful, on average, relative to their control accounts, while brighter, redder, squares represent city-month pairs in which they produced less unrestful tweets than their controls. Pale squares represent city-month pairs in which there was little substantial difference. This figure illustrates the fact some of the cities that had more extreme contrast in the cross-section were actually **not** the ones that drove the jump observed in August, 2016. For example, Richmond was the city with the largest cross-sectional general unrest contrast, but its contrast actually decreased over time. The large jump in August was driven by increases in relative unrest in cities in several swing states, Ohio (Cincinnati & Cleveland), Pennsylvania (Pittsburgh & Philadelphia & Camden, NJ), Arizona (Phoenix), as well as two major media markets: New York and Boston.

But there is no statistically significant relationship between the size of the August jump and whether the city was located in a swing state in the 2016 U.S. presidential election. Finally, this figure makes clear that one of the outliers in cross-sectional results, Minneapolis, is probably driven by its narrow window of time in which it operated, one month early in the campaign.

Discussion

We find strong evidence supporting all three of our hypotheses, with results that are broadly consistent with what is already known about IRA strategy. For H1, the IRA Newsfeeds produced more unrestful tweets than their matched control accounts. This is true both in our hand-coded sample and in our algorithmically classified sample and it is robust to restricting our attention to differences that arise within the month-city cluster. It is also true for the entirely "hands off" WNLU emotion classifier, where the IRA's tweets were laden with markers of negative emotions, like disgust, sadness, and danger, and likely to include markers of joy. This overall pattern is consistent with an attempt to set the political agenda to focus more on unrest than would occur without their intervention. This approach is quite consistent with the IRA's behavior on other accounts, where they cultivated accounts within affinity groups with extreme views (Linvill & Warren, 2020b), but with a subtler mechanism.

For H2, the IRA Newsfeeds differ from the controls particularly with respect to unrest that is human originating. These include labels such as identity danger, crime, and (to a lesser extent) institutional failure. As with overall unrest, this is robustly true across specifications. The WNLU labels are not well suited to address this hypothesis. This pattern, focused on **people,** is also consistent with the behavior on other account types, where the goal seemed to be to divide interested groups against each other (NOT US… someone else.).

For H3, there is strong evidence of substantial change in IRA Newsfeed behavior in August, 2016, when they go from slightly over-emphasizing unrest to strongly over-emphasizing it, particularly for crime and identity danger. Once again, a sudden shift in behavior is completely consistent with the activity by the other account types, where Left and Right Trolls enacted a dramatic shift in their output in the weeks before the election (Linvill & Warren, 2020a).

These results tell a consistent story about how the IRA Newsfeeds (attempt to) use an agenda-setting approach to affect perceptions about the level and mix of unrest and uncertainty in the U.S. in the months before the 2016 U.S. Presidential election. The August change in behavior also undermines possible alternate hypotheses for why the IRA employed more unrestful content. Local news outlets, use violent and unrestful content as a mechanism to reach and grow audiences (Kimberly & Aday, 2003). One could hypothesize this was the IRA's motivation in increasing the percentage of unrestful content, but if this were the case why wait so long? Further, it seems likely that genuine local newsfeeds already take audience growth into account in their specific mix of content and distribute content with the maximization of user engagement in mind. The IRA's divergence from this same mix a few months before the 2016 election, just as their Left Troll and Right Troll account types were changing behavior away from a growth focus (Linvill & Warren, 2020a), points to strategic reasoning.

Although the analysis of other IRA account types motivated predictions about how the agenda-setting strategy would vary over time, we had no strong a priori predictions about how it would vary over geography. Given the political messaging throughout the more explicit account types, and the well-documented relationship to electoral politics, it seemed plausible that greater efforts to skew the agenda might be applied for accounts that targeted cities in likely swing states in the presidential election. There are a substantial number of cities with media markets in swing

states included in the set, but we found no statistically significant evidence of particularly large agenda-setting efforts in these states, either overall or during the August shift. There is substantial heterogeneity, both overall and in the size of the August shift, but the drivers of this heterogeneity remain unclear.

Conclusions

The role that the Newsfeeds played in the IRA's clandestine social media campaign has not been well understood. Various theories have been expounded, but this analysis provides the first explanation for their existence. These results provide another example of how the IRA used specialized accounts (Linvill & Warren, 2020b) in specialized ways to achieve their ends, including influencing the political discourse in ways that polarized. It is well-documented that the ideological types (Right and Left Trolls) invested in infiltrating affinity groups in an apparent attempt to pull those groups apart (Linvill et al., 2019). The Newsfeeds pursued the same end through different means, by establishing a reputation as disinterested purveyors of real local news and then drawing on this established reputation to present a funhouse version of the world in which society was breaking down and more rife with crime and identity-group conflict than in actual fact.

Beyond the IRA, this study illustrates how social media can be systematically utilized for inauthentic agenda setting purposes. It seems possible that the IRA's efforts may not be unique. In the run up to the 2020 U.S. election domestic, partisan actors were identified engaging in similar behavior (Glazer & Hagey, 2020). We don't know of examples of this strategy being applied in other state sponsored, inauthentic social media campaigns, nor do we see examples of it in the Twitter Informational Operations archive.

Most models of an agenda-setting media show media as driven by something like profit maximization. To the extent that an outlet over-emphasizes dramatic or conflictual news, it is

likely often motivated by a drive toward expanding its readership -- “if it bleeds it leads.” This research has identified a different sort of actor that might engage in agenda setting behavior but for different reasons. In this case, as election day neared the IRA did not seem interested in attracting users but rather only influencing their perceptions. Similar actors, therefore, will likely be less affected by the disciplining power of competition (Mullainathan & Shleifer, 2005) in limiting their extremism.

Our findings do not answer, however, why unrestful content was not significantly higher than the controls after August 2016 in six cities. There are many possible explanations. It could be that these cities were used as a control group by the IRA for their own analytics purposes or simply that their method of biasing the content was imperfect. Our findings leave this question open.

The biggest space for future work, here, is to analyze whether this sort of agenda setting can work. We know that real media has the power to shape how readers experience the world. And, in the specific context of economics news, there is some evidence that bias in reporting can bleed over into bias in readers’ beliefs (Lott & Hassett, 2014). It seems natural that that power would extend to these inauthentic purveyors of real news, but we have not demonstrated that here.

References

Arif, A., Stewart, L., & Starbird, K. (2018). Acting the part: Examining information operations within #BlackLivesMater discourse, proceedings of the ACM on Human-Computer Interaction, Nov. doi:10.1145/3274289

Bastos, M., & Farkas, J. (2019). "Donald Trump is my President": The Internet Research Agency propaganda machine. *Social Media + Society,* 1-13. doi:10.1177/2056305119865466

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5,* 135–146. doi:10.1162/tacl_a_00051

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research* . Thousand Oaks, CA: Sage.

DellaVigna, S., & Kaplan, E. (2007). The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics, 122,* 1187-1234. doi:10.1162/qjec.122.3.1187

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, r., Fox, R., Albright, J., Johnson, B. (2019). *The Tactics & Tropes of the Internet Research Agency*. Retrieved from https://disinformationreport.blob.core.windows.net/disinformation-report/NewKnowledge-Disinformation-Report-Whitepaper.pdf

Dreier, P., & Martin, C. R. (2010). How ACORN was framed: Political controversy and media agenda setting. *Perspectives on Politics, 8,* 761-792. doi:10.1017/S1537592710002069

Erbring, L., Goldenberg, E. N., & Miller, A. H. (1980). Front-page news and real-world cues: A new look at agenda-setting by the media. *American Journal of Political Science, 24,* 16-49. doi:10.2307/2110923

Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y., & Adams, K. (2020). Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review.* doi:10.1177/0894439320914853

Ghanem, S. (1997). Filling in the tapestry: The second level of agenda-setting. In M. McCombs, D. L. Shaw, & D. Weaver (Eds.), Communication and democracy (pp. 3–14). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Glazer, E., & Hagey, K. (2020, October 19). Partisan sites posing as local news expand ahead of election. *The Wall Street Journal.* Retrieved from https://www.wsj.com/articles/partisan-sites-posing-as-local-news-expand-ahead-of-election-11603077119

Gross, K., & Aday S. (2003). The scary world in your living room and neighborhood: Using local news broadcast news, neighborhood crime rates, and personal experience to test agenda setting and cultivation. *Journal of Communication, 53,* 411-426. doi:10.1111/j.1460-2466.2003.tb02599.x

Iyengar, S., Kinder, D. R., & Peters, M. D. (1982). Experimental demonstrations of the "not-so-minimal" consequences of television news programs. *The American Political Science Review, 76,* 848-858. doi:10.2307/1962976

Jamieson, K. H. (2018). Cyber-War: How Russian hackers and trolls helped Elect a president. New York: Oxford University Press.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv:1607.01759 [cs.CL]

Keller, F., Schoch, D., Stier, S., & Yang, J.H. (2020) Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication, 37,* 256-280, DOI: 10.1080/10584609.2019.1661888

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review, 111,* 484-501. doi:10.1017/S0003055417000144

Kiousis, S., Mitrook, M., Wu, X., & Seltzer, T. (2006) First- and second-level agenda-building and agenda-setting effects: Exploring the linkages among candidate news releases, media coverage, and public opinion during the 2002 Florida gubernatorial election. *Journal of Public Relations Research, 18,* 265-285. doi:10.1207/s1532754xjprr1803_4

Linvill, D. L., Boatwright, B., Grant, W., & Warren, P. L. (2019). "The Russians are hacking my brain!' investigating Russia's Internet Research Agency Twitter tactics during the 2016 United States presidential campaign. *Computers in Human Behavior, 99,* 292–300. doi:10.1016/j.chb.2019.05.027

Linvill D. L. and Warren, P. L. (2020a). Engaging with others: How the IRA coordinated information operation made friends. *Harvard Kennedy School Misinformation Review.* Retrieved from https://misinforeview.hks.harvard.edu/article/engaging-ira-coordinated-information-operation/

Linvill D. L. and Warren, P. L. (2020b) Troll factories: Manufacturing specialized disinformation on Twitter, *Political Communication, 37,* 447-467. doi:10.1080/10584609.2020.1718257

Lott, J., & Hassett, K. (2014). Is newspaper coverage of economic events politically biased? *Public Choice,* 160(1/2), 65-108.

Lumley, T. (2020). biglm: bounded memory linear and generalized linear models. R package version 0.9-2. Retrieved from https://CRAN.R-project.org/package=biglm

McCombs, M. (2005) A Look at agenda-setting: Past, present and future. *Journalism Studies, 6,* 543-557, doi:10.1080/14616700500250438

McCombs, M. (1994). News influence on our pictures of the world. In J. Bryant & D Zillmann (Eds.), Media effects: Advances in theory and research (pp. 1–16). Hillsdale, NJ: Erlbaum.

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly, 36,* 176–187. doi:10.1086/267990

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111-3119).

Moran, M. (2003). Arguments for rejecting the sequential Bonferroni in ecological studies. *Oikos, 100,* 403–405. doi:10.1034/j.1600-0706.2003.12010.x

Mullainathan, Sendhil, and Andrei Shleifer. 2005. "The Market for News." American Economic Review, 95 (4): 1031-1053. doi:10.1257/0002828054825619

Pampel, F. C. (2000). *Logistic Regression: A primer.* Thousand Oaks, CA: Sage.

Pierri, F., Artoni, A., Ceri, S. (2020). Investigating Italian disinformation on Twitter in the context of 2019 European elections. *PLoS ONE, 15.* doi:10.1371/journal.pone.0227821

Roberts, M. (2018) *Censored: Distraction and Diversion Inside China's Great Firewall.* Princeton University Press, Princeton, NJ.

Tsay, B. (2013). bigcluster. GitHub repository. Retrieved from https://github.com/brtsay/bigcluster

Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society, 20*, 2028-2049. doi:10.1177/1461444817712086

Figure 1

Sample IRA newsfeed accounts

Figure 2

ROC curves for the classifiers trained on the hand-labelled data, evaluated on a holdout set

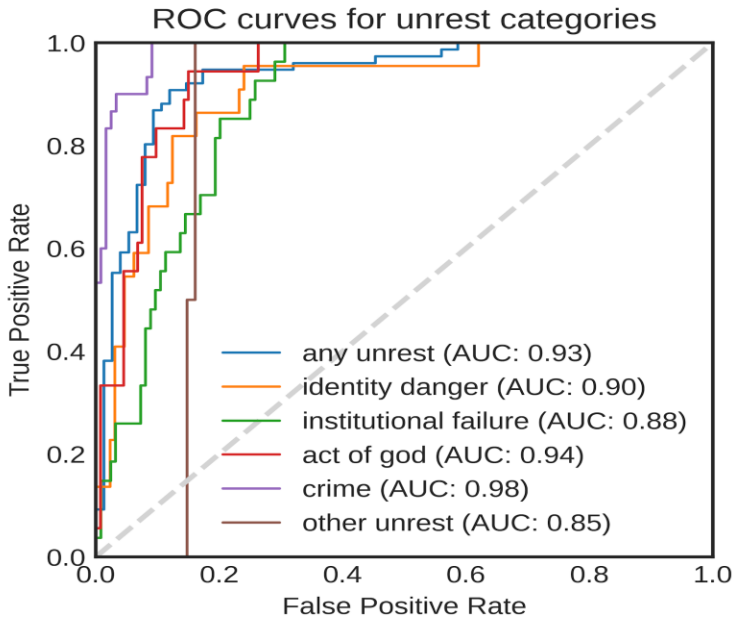


Figure 3

Precision-recall curves for the classifiers trained on the hand-labelled data, evaluated on a holdout set

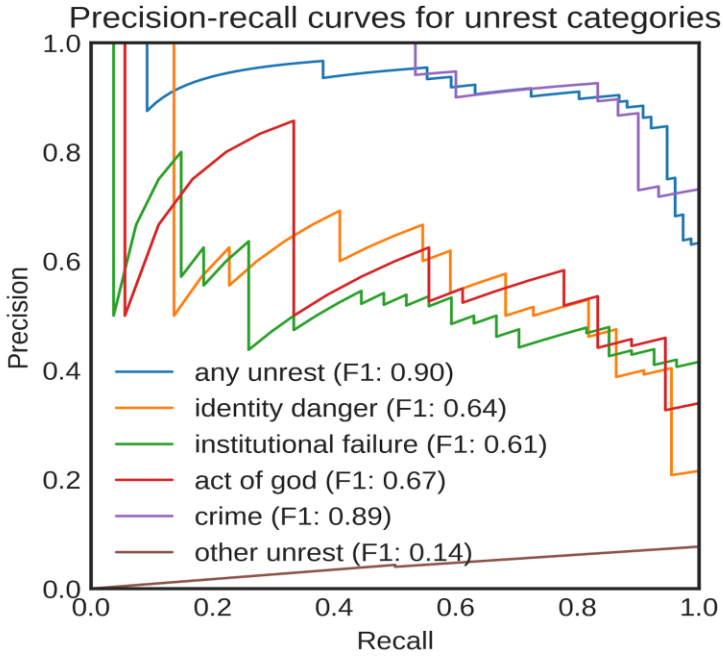


Figure 4

Estimated Coefficient on IRA Newsfeed Indicator and 90% Confidence Intervals by Type of Unrest and City Targeted by Troll Newsfeed

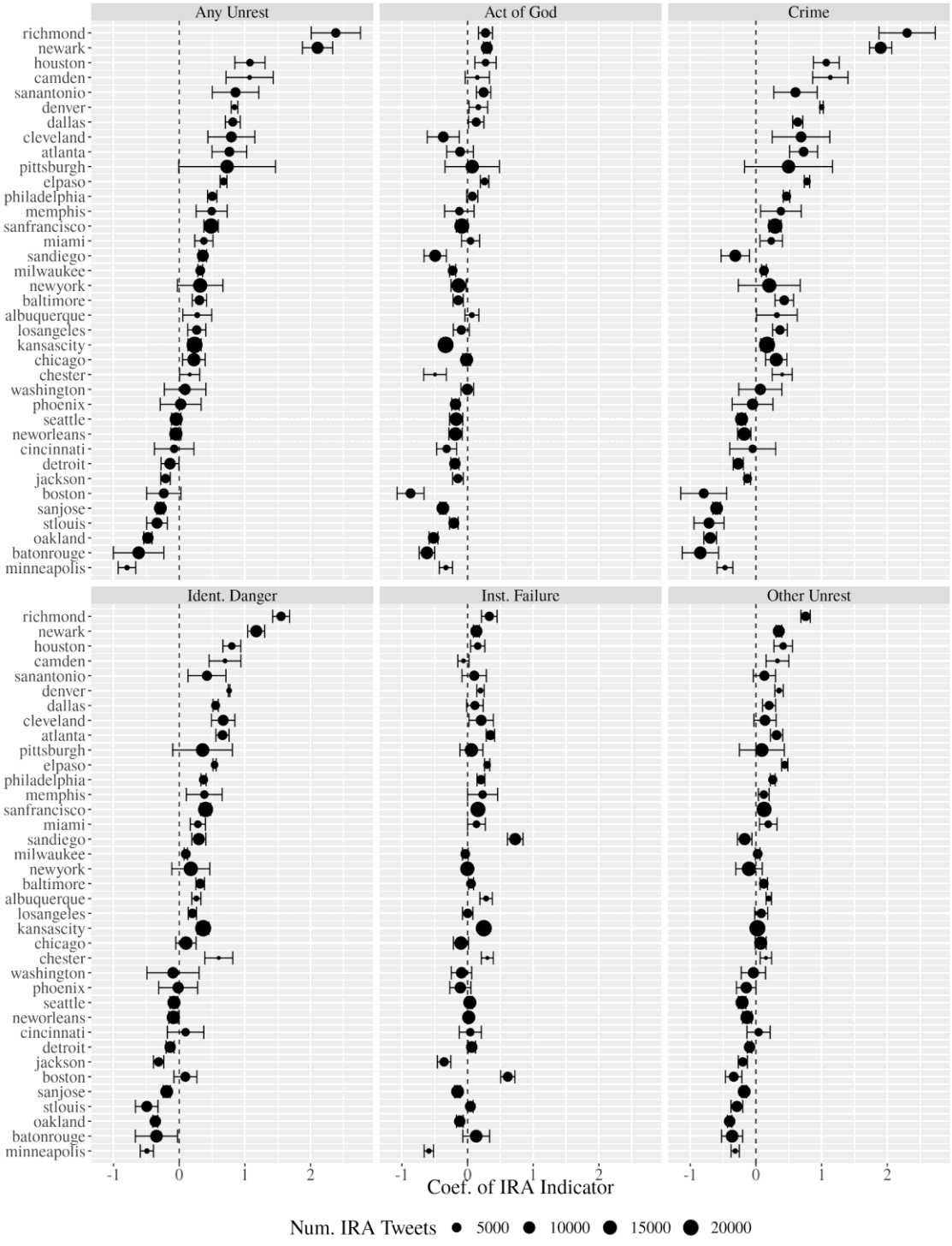


Figure 5

Estimated Coefficient on IRA Newsfeed indicator and 90% Confidence Intervals by Type of Unrest and Month

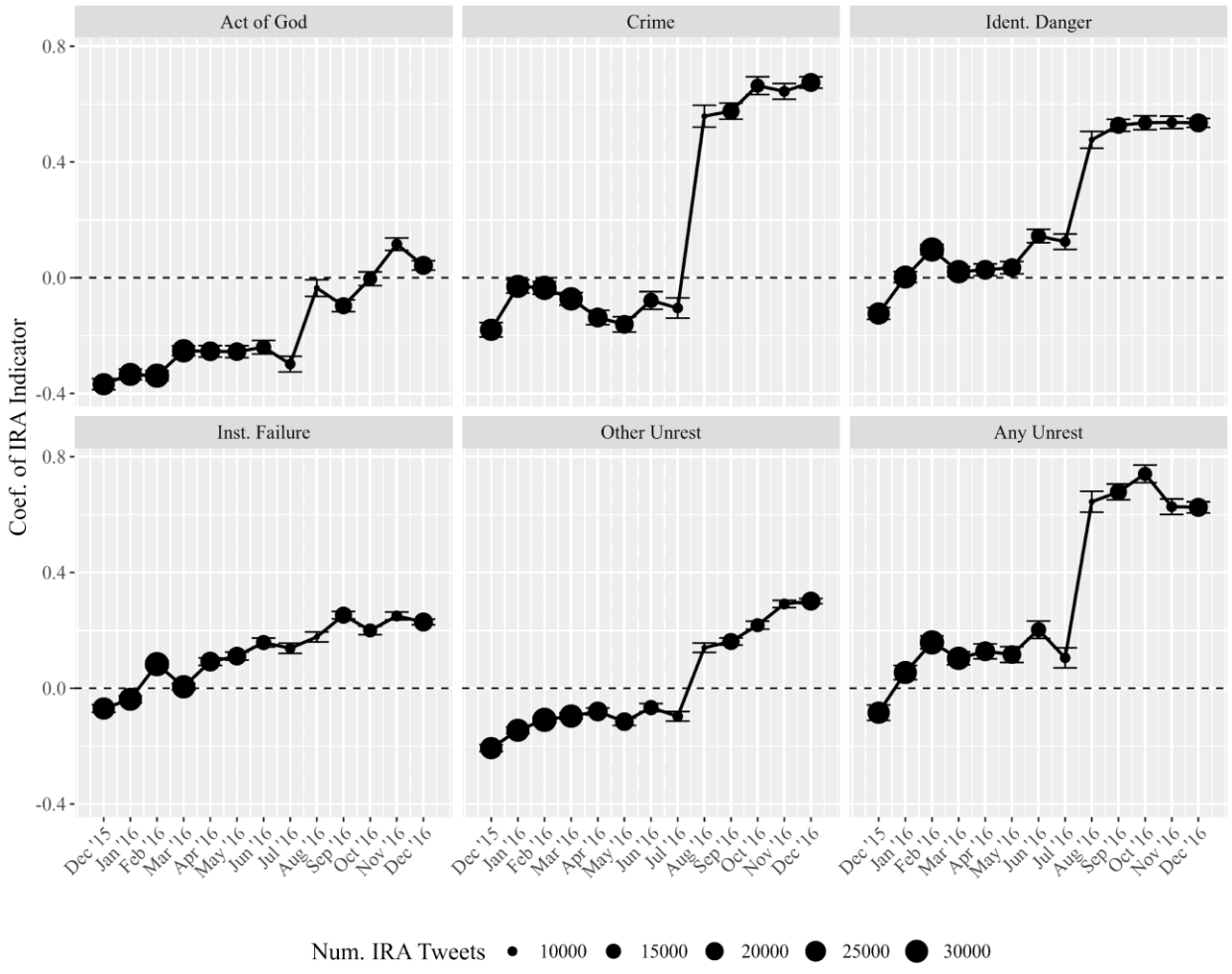


Figure 6

Estimated Coefficient on IRA Newsfeed Indicator by Type of Unrest and City-Month

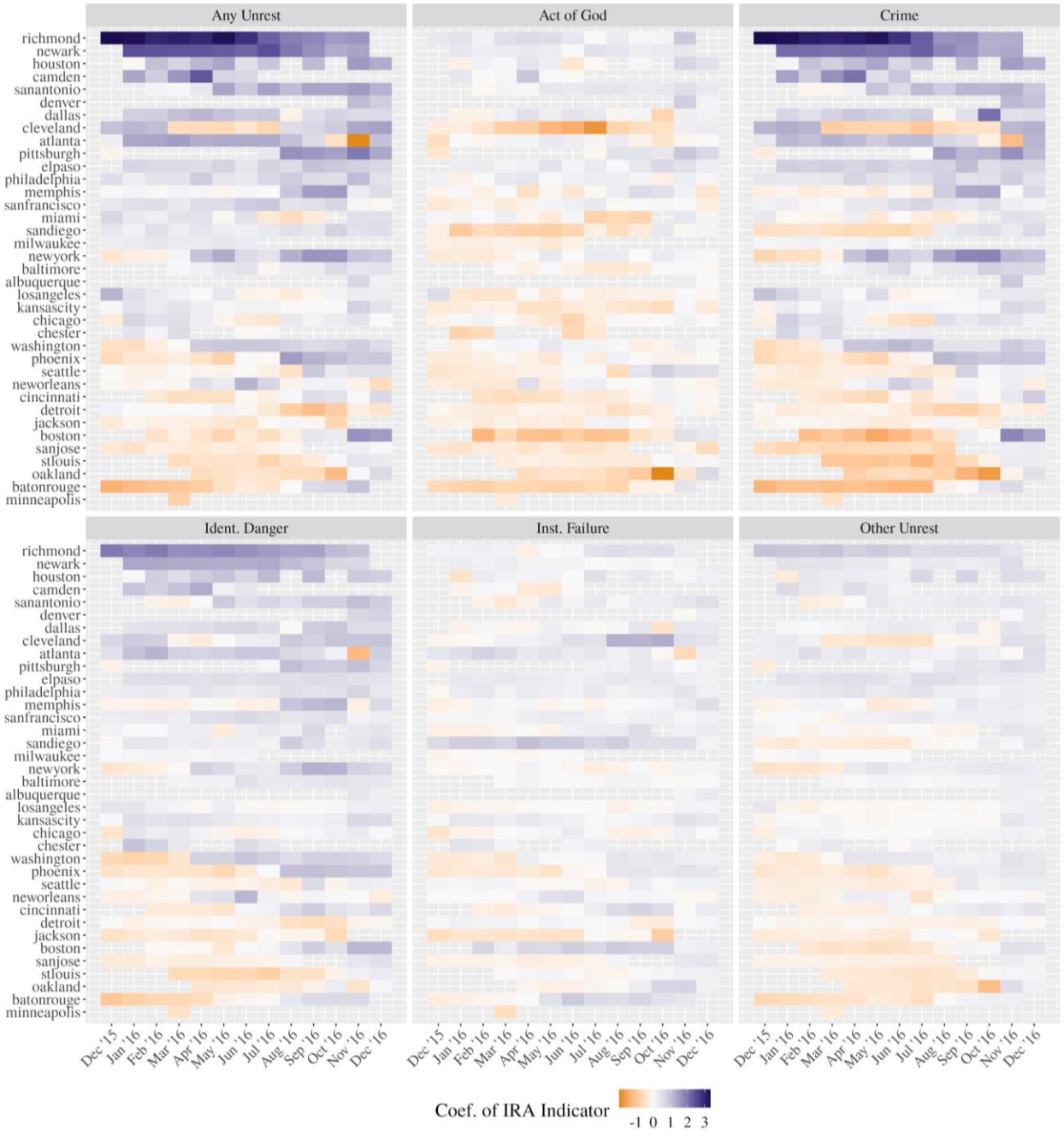


Table 1

Examples of tweets and classifier output for various label categories. The bottom two rows show examples of (infrequent) incorrect model output

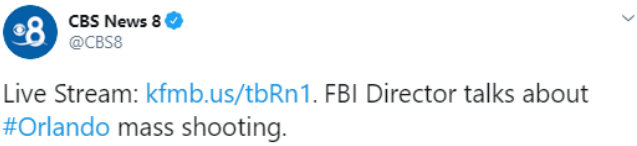
Tweet	Unrest category and confidence	Result evaluation
	Crime: 0.94	Correct (true positive)
	Institutional failure: 0.02	Correct (true negative)
	Identity danger: 0.07	Incorrect (false negative)
	Act of god: 0.92	Incorrect (false positive)

Table 2

Mean probabilities of category membership (estimated with the logistic regression classifier) for tweets in the full data set for IRA and control tweets both before and after August 2016.

Label	Mean probability for control tweets prior to August 2016	Mean probability for IRA tweets prior to August 2016	Mean probability for control tweets from August 2016 onwards	Mean probability for IRA tweets from August 2016 onwards
Any unrest	.549	.584	.483	.624
Identity danger	.123	.130	.105	.159
Institutional failure	.123	.128	.115	.138
Act of god	.061	.047	.062	.062
Crime	.104	.096	.083	.144

Table 3

Regression coefficients and corresponding Bonferroni-corrected p-values for the "is IRA" feature in each of the five unrest category analyses.*

Dependent var. (logit transformed)	Regression coefficient	Exp. reg. coef.	p-value
Any unrest	0.292	1.34	3.8e-12
Identity danger	0.215	1.24	1.9e-10
Institutional failure	0.105	1.11	7.2e-08
Act of god	-0.174	0.84	1.3e-17
Crime	0.17	1.19	9.7e-4

* *The column "Exp. reg. coef." gives the estimated ratio of the odds of an IRA tweet being in the relevant category to the odds of a non-IRA tweet being in that category.*

Table 4

Regression coefficients and corresponding Bonferroni-corrected p-values for the "is IRA" feature in each of the six WNLU sentiment and emotion analyses.*

Dependent var. (logit transformed)	Regression coefficient	Exp. reg. coef.	p-value
Anger	0.199	1.22	3.6e-32
Disgust	0.410	1.51	1.0e-64
Fear	0.064	1.07	4.3e-3
Joy	-0.0129	0.88	2.7e-5
Sadness	0.248	1.28	7.7e-32
Sentiment	-0.0332	0.97	0.22

* The column "Exp. reg. coef." gives the estimated ratio of the odds of an IRA tweet being in the relevant category to the odds of a non-IRA tweet being in that category.