Violence against journalists and freedom of the press: Evidence from Mexico*

Andres Jurado
Brown University

Juan S. Morales
Collegio Carlo Alberto
University of Turin

July 6, 2020

Abstract

This paper studies how the murder of journalists affects press coverage in Mexico. We use data on news reporting collected from a comprehensive archive of the largest Mexican news outlets and a dataset of over 6 million *tweets* published by Mexican journalists and media outlets. We find large reductions in intensity of reporting among victimized outlets following an aggression. This reduction is consistent with censorship, as the public is interested in these events and engages more with their content following an attack. These attacks affect *how* the press reports the news as well: we measure subtle, but persistent, changes in tone of coverage, with outlets that remain active emphasizing the most violent aspects of organized crime post-attack. We also document indirect and localized spillover effects on non-victimized journalists using a triple-difference design, in particular, after the killings we observe reductions in reporting for journalists who followed the victimized outlet on Twitter and were located in the same state. Finally, using census data we show that states with higher levels of violence against journalists between 2010 and 2015 saw reductions in the number of active journalists and changes in their demographic composition.

*We thank seminar and conference participants at Brown University, the Collegio Carlo Alberto and the 2019 HiCN Workshop (Paris School of Economics) for helpful comments and discussions. All errors are our own. jose_jurado_vadillo@brown.edu, juan.morales@carloalberto.org.

1 Introduction

Freedom of the press is regarded as one of the pillars of democratic institutions. Over the last decade, more than 550 journalists were killed around the world, threatening this freedom and weakening democracy.[1] This paper studies the relationship between targeted media violence and news reporting in Mexico, a country suffering from high levels of drug-trafficking related violence and one the highest murder rates in the world, including five percent of all journalists' killings.

Despite the frequency of targeted violence against the media, systematic evidence examining its effects on news reporting remains scarce. Qualitative evidence suggests that harassment and violence against the media from criminal organizations could have mixed effects on news reporting. First, violence may *deter* the press from covering certain events. A journalist working along the US border reports that:

> "We still haven't shaken the fear that we had at one point, that's to say there are many things that could be investigated but that aren't."
> (Relly and González de Bustamante, 2014, p. 115)

On the other hand, threats and violence may also cause journalists' *backlash*:

> "If they call us to tell us what to do, or what not to publish, we're going to publish it twice over and we're also going to write that they called us to tell us not to publish."
> (Relly and González de Bustamante, 2014, p. 116)

We study the relationship between violence against the press and news reporting using data on all reported media workers' killings in Mexico between 2006 and 2017. We combine these reports with newly compiled data of news articles and press activity from both traditional and online media sources, including Twitter and *Eficiencia Informativa* — a private database with over 35 million news articles from Mexican news outlets. Our Twitter dataset consists of around 7 million tweets published by more than 300 media user

[1] Committee to Protect Journalists, 2019 (https://cpj.org/), accessed April 9, 2020.

accounts between 2009 and 2017. These include 76 outlets with at least one of their employees killed, and 226 popular Mexican journalists. In addition, we use data from the Mexican census to study long-term changes in the demographics within the profession.

We exploit the precise timing of journalists' killings in a series of event-study empirical exercises, through which we measure short and medium-run changes in news activity following the violent incidents. Importantly, we differentiate between the *direct* effects of being a victim of targeted violence, and the potential *indirect* effects that may spillover onto non-victimized outlets and journalists.

We observe sustained reductions of around 25% in Twitter activity by victimized outlets following the murder of a media worker. Smaller outlets are especially sensible to these attacks, with 20% of them exiting the market for crime news permanently. We also document *indirect* effects that are highly localized and that are concentrated among journalist likely more at risk, i.e., those in the same Twitter network and state as the victim.

We use a modern supervised machine learning algorithm to quantify changes in tone in Twitter among victimized outlets following an attack, and find persistent effects. *How* tone changed is harder to characterize. We show that words that emphasize the most violent aspect of organized crime predict post-attack, whereas prior to it we find mostly words that describe law-enforcement operations. We also measure a change in polarity towards more *negative* speech afterwards.

Since 2010, the Census and the Inter-Census Survey has included new questions that allow us to identify the number of journalists active in the 32 states in the country. We implement a difference-in-difference estimator to measure long-term effects of violence against the press. States with more homicides have fewer journalists than predicted. Demographics also appear to change as a results of these attacks with journalists being less likely to be married and have kids, and earning lower wages.

Our work is related to the literature on the political economy of media that examines both the determinants and the effects of media coverage. The importance of media for

political processes has been widely established. For instance, Adena et al. (2015) show that radio was important in the rise of the Nazi party in Germany, Snyder and Stroemberg (2010) document that the US congress is more accountable when the press covers local issues, and Dellavigna and Kaplan (2007) show an effect of exposure to Fox News on voting for the Republican party. Nevertheless, there is also evidence on limits to the influence of news media: consumers tend to discount information from sources that are considered biased (Chiang and Knight, 2011), and some behavioral mechanisms that can limit the influence of the press include switching and tuning out (Durante and Knight, 2012; Knight and Tribin, 2019).

Closer to our work, several papers have studied how the media can be influenced or biased: di Tella and Franceschelli (2011) show that public funds spent in advertising affect how the press covers negative events involving the government, while Beattie, Durante and Knight (2017) and Reuter and Zitzewitz (2006) show that private financial incentives can change how the press covers the news. Importantly, Gentzkow and Shapiro (2010) document that newspapers in the US respond to public preference for like-minded news by choosing slant in a profit-maximizing manner. In the context of Mexico, Ramírez Alvarez (2017) reveals how editorial standards and informal agreements between publishing houses and the government influenced how the press covered the war on drugs.

Outside of economics, a number of papers have studied repression of journalists in Mexico, most notably in the fields of political science and journalism. Stanig (2015) documents a negative relationship between media regulation, in the form of defamation laws, and coverage of corruption cases across Mexican states. Studying the determinants of attacks against journalists, Holland and Rios (2017) show that rivalries between cartels are a significant predictor of targeted media violence, while Hughes and Márquez-Ramírez (2018) document that being in an environment with higher criminal violence is positively associated with the incidence of threats against journalists. Most recently, Salazar (2019) shows that, though aggression against journalists reduces the number of headlines criti-

cal of the government, the presence of newspaper networks and NGOs can mitigate these negative effects.

2 Context

Between January 2006 and October 15, 2017, the Committee to Protect Journalists (CPJ) documented the murder of 104 media workers in Mexico, making it one of the deadliest countries for journalists worldwide. These homicides are highly targeted operations: victims covered crime in a majority of cases and the *modus operandi* appears consistent with that of organized crime. Out of the 104 documented killings, 11 journalists were kidnapped before being murdered, and 9 were tortured in addition. Of the killings, 84% were deemed by CPJ as "murders", while the rest were attributed to "dangerous assignments". In 79% of cases the CPJ attributed the attack to criminal groups, followed by government and the military (7.9% each), and to local residents (5.3%). At least 22 journalists received threats (*ibid*).

Organized crime often tries to influence news reporting, suggesting that these murders ultimately tried to punish individuals or outlets that were "out of line":

> "In the Spring of 2010, it came to our attention that there was a spokesman for organized crime. In the coming days a reporter –on behalf of this individual– scheduled a meeting with a group of colleagues. They warned us about who called the meeting and what would happen if we didn't attend [...] The spokesman explained the new rules: no one publishes material without approval of the "boss"; no one is allowed to ignore phone calls from them; no one can refuse to accept bribes [...]"[2]
>
> (Valdez, 2016, p. 42)

Figure 1 presents the proportion of victims that reported on various popular subjects, based on CPJ detailed accounts. 78% covered crime or the police beat, 34% politics, and 20% corruption (note that a journalist may cover multiple subjects and hence the columns do not add up to 100%).

[2]Own translation.

The location where these murders occurred also indicates connections to organized crime: states that are important points of entry for drugs tend to also have a higher number of murdered media workers. Figure A1 depicts murders of journalists and homicides among the general population for the 32 states in the country. Journalists are especially at risk in the states of Guerrero, Oaxaca and Veracruz (dark blue), which together account for 40% of all murders, while representing 12.8% of the country's population. Guerrero and Veracruz are notorious for being known ports of entry for cocaine imported from Colombia, and other illegal substances, while Oaxaca suffers occasional outbreaks of ethnic tensions. The eastern state of Veracruz is by far the most hostile for the press, with 18% of all victims. Attacks on members of the press are not a simple function of overall violence, as can be seen by the weak correlation between total number of homicides and murders of members of the press in the country. Northern states, especially those at the border with the US, experience high homicide rates, but report comparatively few journalists killed.

These levels of violence are a recent phenomenon and are linked to the start of the period commonly referred to as the Mexican War on Drugs. Figure A2 shows how the targeting of media workers increased steeply by 2006, which was followed two years later by more homicides among the general population. 2006 marks the last year of the inter-cartel war between the criminal syndicate the *Federation* and some of the Gulf-based cartels. With a cease-fire in place in 2007 we observe both a reduction in number of media workers murdered, as well as in total homicides in the country. By 2008, Pres. Calderón militarized strategy against DTO's was underway and the *Federation* splintered into two bands fighting for control of profitable drug-trafficking routes. This coincides with the increase in homicides among media workers and the general population that persists today.

Perhaps one reason behind this high level of victimization is the few number of cases that are solved by the police. CPJ reports that in 86.8% of cases there was complete impunity, and in 10.5% partial impunity. Justice was delivered only in 2.7% of cases. In a 2017

report it concludes that[3]:

> Endemic impunity allows criminal gangs, corrupt officials, and cartels to silence their critics. [...] Despite federal government efforts to combat this deadly cycle, justice remains elusive, and impunity the norm.

3 Data

3.1 Attacks on journalists

Article 19 and the Commitee to Protect Journalists (CPJ) are two leading NGO's that advocate for journalists. Both keep track of media workers murdered in the country, along with known affiliations. Out of 104 victims, 17 were classified as free-lancers and hence are not matched to any outlet. The rest is assigned to 103 outlets, as some journalists have more than one affiliation.

Both of these organization report the place and date of death. Sometimes the date of death is uncertain if for instance the media worker was first kidnapped and his or her body was later found. In those cases we follow the convention of considering the earliest plausible date.

3.2 "National" outlets

Eficiencia Informativa (**EFIC**) is a private company that collects data in real time from printed and electronic media, predominantly from Mexican outlets. Their archive contains over 35 million full news articles, tweets, radio and TV transcripts. The company is also certified ISO 9000 for quality management.

We identified a sample of 2.2 million news articles by collecting items that match any of the following keywords: "narc*", "sicari*" (hitman), "difund*" (spread as in "spread infor-

[3]Washington Post, *The Most Common Punishment for Killing a Journalist in Mexico: Nothing*, https://www.washingtonpost.com/news/worldviews/wp/2017/05/03/the-most-common-punishment-for-killing-a-journalist-in-mexico-nothing/. Accessed March 19, 2020.

mation"), "crim*", "PGR" (federal attorney), "enfrentamiento" (confrontation), "ejército" (army), "drogas" (illegal drugs), "fosas" (illicit graves), "ejecutado" (extra-judicial execution). This list is based on terms with the highest coefficients from a LASSO regression from Ramírez Alvarez (2017), which predicts whether an article is related to drug-trafficking or not. We have included in our list additional terms that we believe also describe the Mexican War on Drugs and other crime-related events. These articles range from January 2006 to October 15, 2017.

22.5% of items in this dataset are transcripts of radio broadcasts, 16.5% of TV, 0.2% correspond to the Official Gazette of the Federation (*Diario Oficial de la Federación*). The rest, 60.8%, is printed media. EFIC tends to track outlets with high circulation, as these are more likely to be of interest to its target audience: government, private individuals and corporations. While our database contains 169 outlets, two leading newspapers, *La Jornada* and *Reforma*, account for 248 thousand articles or 11.2% of the total.

Our sample from EFIC contains a large number of terms related to drug violence, as can be seen in Figure A3, with words such as "security", "federal_attorney", "investigate", "police", "organization", "seize", "dead", "army", "crime", "law". Figure A4 panel *a* plots the monthly number of articles in EFIC to total homicides in the country. These series exhibit a .82 correlation[4], with approximately 10 articles per homicide. The EFIC database, thus, seems to deliver high-quality data to investigate the questions of interest.

3.3 Twitter

We built two distinct datasets of tweets using a selected set of usernames. The first dataset includes tweets for 224 journalists whose usernames were collected from twitter-mexico.com, a website which archived and documented popular Twitter users in Mexico.[5] The second set of usernames belongs to 76 victimized outlets, as documented by the CPJ and by Article

[4]The spike in December 2014 is likely due to extensive coverage of the massacre of 43 students in Ayotzinapa, an event that generated headlines worldwide. See for example this article from the New York Times.

[5]Though the site is no longer online, the journalists page we used can be accessed through the wayback machine.

19, which we were able to manually match to a corresponding Twitter account.

We then used Twitter's *Advanced Search* tools and the Twitter API to collect tweets published by the selected users between 2009 and 2017. One important limitation of the collection methodology is that only the last 20 tweets per user per day can be retrieved. Since often users do not tweet more than twenty times in one day, this limit does not always bind.[6] Our final datasets comprise 5 million tweets published by Mexican journalists, and 1.8 million tweets published by victimized outlets. At the username-day level, the 20-tweet limit binds for 17 percent of observations for the journalists dataset, and for 42 percent of observations for the outlets dataset.

We identify news about crime using a broad set of keywords related to drug-trafficking, violence and corruption. We classify these as *violence tweets*, and they make up almost 6 percent of tweets in the journalists dataset, and around 14 percent of tweets in the outlet database.[7] Figure A4 panel *b* shows that tweets with violent content by the top journalists in the country have a .43 correlation with the monthly number of homicides in the country. This correlation is likely underestimated due to censoring, and an almost 5-fold increase in tweets in December 2014, which was likely caused by coverage of the massacre of Ayotzinapa (see footnote 4).

3.4 Other data

We gathered data on daily circulation figures for 35 victimized outlets from the State Secretariat's National Census of Printed Media[8] (*Padrón Nacional de Medios Impresos*). The Census records the municipalities where outlets are distributed, as well. Information is collected

[6]The limitation of the data collection methodology results from the historical nature of the data. One of our main outcomes (number of published tweets) will be measured with error, biasing our coefficients towards zero, such that our estimates can be considered conservative estimates of the true effect. We discuss this in more detail below. See Morales (2020) for more details on the data collection methodology.

[7]The set of words is: *cartel*, *narco*, violence, homicide, death, body, threat, justice, alleged, accuse, criminal, assassin, kidnap, forced disappearance, victim, convict, drug, government, corrupt, police, military, general attorney, torture, conflict, war, *Chapo*, investigation, impunity, crime, ties to, arrest, member of, confrontation, injured.

[8]Compiled through April 2020. https://pnmi.segob.gob.mx.

by third-party auditors, which might be onerous for smaller outlets. Hence, larger outlets tend to be over-represented.

General population homicides are reported by the National Statistics Institute (*INEGI*). For the present article we use daily counts of homicides at the municipality and state level. CIDE's Drug Policy Program (PPD) maintains a leaked database with homicides attributed by a government panel to drug trafficking organizations. We decided against using it, as the Mexican government ceased to keep track of murders of this kind in 2011.

Finally, we access data on the 2010 and 2015 Mexican censuses through IPUMS international.[9] The sample of analysis is restricted to journalists and occupational categories in the census that are near that of journalists, which include: accountants, researchers, psychologists, artists and performers. Summary statistics for the journalists in our sample are in table A1, separately across states: those where at least one journalist was murdered, and those with no journalists' murders.

4 Empirical Analysis

Different mechanisms, acting in opposite directions, might shape the response to an attack. We characterize these mechanisms adhering to previous work when appropriate (Pan and Siegel, 2019). Volume of coverage might decrease simply because there are fewer journalists reporting for a given outlet, which we refer to as a *mechanical* effect. The pool of news to report includes now a high profile murder (*content* effect). In addition, outlets may change their publishing *behavior* in response to the attack. The objective of the criminals is likely to *deter* future publishing through fear. Outlets may reduce the intensity of reporting or change their tone of coverage to minimize the risk of a subsequent attack.

Behavioral responses might lead to more assertive coverage, however: the press may step up its publishing to signal that they will not be intimidated, or to protest the murder

[9] We do no use previous censuses as we can not identify journalists as an occupational category before 2010.

of a colleague (*backlash* effect). The attack itself might increase public interest in the content from the targeted outlet, which could lead to more intense publishing activity in turn (*demand* effect). These mechanisms in turn may be borne out both by the targeted outlet (*direct* effects), or by other journalists who observe the attacks (*indirect* effects).

In section 4.2 we show that attacks on journalists reduced the volume of tweets published by targeted outlets, while engagement (likes and re-tweets) and Google searches for homicides of journalists increased substantially (section 4.3). Demand forces and content effects, it appears, are of secondary importance and the combination of mechanical and deterrence results in what we view as censorship. Our baseline specifications control for calendar month interacted with state fixed effects, which account for all time varying unobservables in the states where outlets are located. The effects measured are concentrated among victimized outlets, and spillover effects within the same state appear to be limited in the short run. In the long-run, however, we observe both fewer individuals working as journalists in states that experienced higher levels of victimization, and changes in the composition of the media workforce as revealed by differences in demographic characteristics (section 4.6).

Our ability to distinguish between mechanical and behavioral effects is, however, limited. For instance, the evidence discussed here is consistent with mechanical effects being the main drivers of censorship, but also with behavioral responses, if media workers in outlets that were not victimized believed that their coverage was unlikely to place them in danger, even after accounting for new information. We find suggestive evidence for both effects. In section 4.2 we show that smaller outlets exited the market for news at greater rates following an attack, which would be consistent with mechanical effects. In section 4.5 we show that journalists most at risk scaled down their Twitter activity compared to those that we estimate were less at risk, which is further evidence of indirect deterrence. Finally, in section 4.4 we discuss how text content became more negative post-aggression, with words describing the most violent aspects of organized crime appearing more often, which

is in line with backlash.

4.1 Overall violence and attacks on the press

We use the timing of attacks against the press to identify the effect of a murder on Twitter activity. Two important challenges to identification are the existence of other newsworthy events that may affect coverage directly by changing the pool of news to report, and the expectation of an attack that may lead to a reduction in reporting on sensitive topics *prior* to the killing. The latter would likely lead us to underestimate any reduction in coverage.

We test whether there is an unusual number of homicides among the general population[10] leading to the aggression. Murders are newsworthy events, and constitute for many outlets one of the sections that most drives sales. Second, an increasing number of murders may indicate increased competition among criminal groups which could make the job of the press riskier, as these organizations try to influence coverage.

We thus first consider event-studies examining whether general population murders, in the state and municipality where the attack occurred, vary with the timing of the events. The model we estimate is given by:

$$\begin{aligned} homicides_{set} &= \gamma_{se} + \gamma_t + \sum_{k=-6}^{6} \beta_k \times monthsSinceKilling_{set} \\ &+ \beta_{pre} \times Pre_{set} + \beta_{post} \times Post_{set} + \epsilon_{set} \end{aligned} \tag{1}$$

where $homicides_{set}$ is the log of homicides (net of those from the press) around event e, in state s, at time t. We include event and state fixed effects γ_{se} and calendar month fixed effects γ_t. The event-study indicators $monthsSinceKilling_{set}$ count 30-day windows since the events occur, and Pre_{set} and $Post_{set}$ are binary variables equal to one for $t < 6$ and $t > 6$,

[10] We consider murders *net* of homicides of journalists. From approximately 33,000 annual homicides in the country in recent years only a dozen correspond to journalists, on average.

respectively.[11] Many states experience more than one homicide of a media worker, in those cases we pair each monthly window to the closest event and define time windows with respect to it. The coefficients of interest β_k are normalized with respect to the event-time window before the event β_{-1}. Standard errors are clustered at the outlet (username) level.

Figure A5 presents the β coefficient estimates from model 1 considering the state where the homicide *occurred* and the municipality where targeted outlets are *located*. States that suffered an aggression against the press reported an increase of 2.5% in homicides compared to the preceding month, but this difference is not statistically significant. The municipalities where victimized outlets are *located* report virtually the same number of homicides. Thus, we do not find strong evidence that the pool of news to report on changed around the dates of the attacks, and we cannot reject that journalists affiliated to victimized outlets did not anticipate an attack. CPJ for instance reports known threats to only 16 victims out of 104, and some appear to have disregarded the threat.[12]

4.2 Direct effects of violence on volume of coverage

We study the effect of an attack on volume of Twitter activity by estimating the following event-study regression:

$$\begin{aligned} y_{msot} &= \gamma_o + \gamma_{st} + \sum_{k=-6}^{12} \beta_k \times monthsSinceKilling_{ot} \\ &+ \delta \times x_{msot} + \beta_{pre} \times Pre_{ot} + \beta_{post} \times Post_{ot} + \epsilon_{msot} \end{aligned} \tag{2}$$

y_{msot} is the log of tweets published by outlet o, located in municipality m in state s,

[11] Time windows are defined based on the day of the attack, such that attacks take place during the first day of time window 0. We choose 12-month windows because states where the media is victimized experience clustering of homicides in time limiting our ability to identify coefficients far from the event date. A stark example is the state of Veracruz where we observe 3 consecutive months with an aggression.

[12] "[Maximino] Rodríguez had received other threats in the past, he said in a December 6, 2016, interview with the news website Culco, adding that he was not afraid to continue his work." https://cpj.org/data/people/maximino-rodriguez/ (accessed July 6, 2020.)

at time t. We include the log of homicides that occurred in the municipality where the outlet is located in the past 30 days as a control, x_{mt}, as it could be correlated with both the likelihood of a journalist being killed and news coverage (though results are similar without this control). Our baseline estimates include state $\times$ month fixed effects γ_{st} and outlet fixed effects γ_o. The coefficients of interest β_k capture the change in Twitter activity for victimized outlets relative to other outlets in the same state (but which were victimized at a different time).

Figure 2 shows our estimates. The volume of tweets decreases by around 25 percent for the full sample and 10 percent for the sample with violence tweets, respectively, although in the short run only the coefficients for the full sample are statistically significant. Publishing falls to its lowest point around 3 to 4 months after the aggression, which could be explained by initial reporting about the attack offsetting a reduction in activity.

One potential concern with our main estimate is whether another event that affects volume of coverage co-occurs with the homicide of a media worker. We report regressions where we control for calendar month and *municipality* fixed effects in figure A8. This accounts for all local events at the municipality level, such as local elections, that may affect our results. Reassuringly, coefficient estimates are similar. More concentrated fixed effects control for local unobserved factors but reduce our sample size to only states and municipalities with more than one victimized outlet.[13] Estimates including calendar month FE (as opposed to state $\times$ calendar month FE) appear also in line with the baseline results, which suggests that indirect spillover effects among non-victimized outlets within the same state may be limited.

We have documented how violence against the media led to reductions in *Twitter* activity. While our primary interest is how *coverage of news* changed, we do not have access to the complete set of newspaper articles from these outlets. In the Appendix we show, however, that our results are similar if we consider a smaller set of articles that we retrieved

[13] In the specification with calendar month and municipality fixed effects we are forced to drop 60% of outlets in our sample.

using Google Custom Search Engine (CSE). Therefore, results using Twitter are likely a good proxy for the effect on newspaper articles.

We investigate heterogeneity in the effects by outlet size. Smaller outlets being disproportionately affected by the attacks may indicate the presence of mechanical effects: losing a reporter could hinder the reporting ability of smaller outlets relatively more than it does for larger outlets. Unfortunately, we do not have access to the number of employees of news outlets. While circulation figures might be a good measure of size, we have this for only a handful of victimized outlets from the State Secretariat's census on media. We consider Twitter engagement, readily available, as a proxy for circulation. Using the outlets for which we have circulation, figure A16 reveals that this may indeed be a good proxy.

A natural way to frame this question is in terms of survival rates for outlets following an attack. We consider outlets with at least one violent tweet in the 30 days preceding an aggression and compute the probability that outlet o tweeted at least once from month j through 12 following an attack, $Pr(\sum_{k=j}^{12} tweets_{ok} > 0), j \in [1, \cdots, 12]$.

Panel *a* of figure A17 shows that in the month after an attack 5% of outlets ceased to be active on Twitter, and 12% no longer tweeted violent news. The number of active outlets drops by another 7% between months 2 and 3. 12 months after an attack, 35% of outlets were no longer active, and 40% no longer tweeted violent news. Panel *b* further breaks down survival rates for low and high engagement outlets, which we define based on the median likes and re-tweets 6 months before an aggression. While initially the same percentage of low and high engagement outlets survive, 15% low engagement outlets (LEO) exit the market between months 2 and 3. LEO's have lower survival rates up until 8 months after the attack, as high engagement outlets (HEO) continue leaving Twitter. Panel *c* shows that while 5% of HEO ceased to tweet violent news one month after an attack, the equivalent figure for LEO's is closer to 20%. Another 10% of LEO's ceases tweeting from month 2 to 3. Thus, smaller outlets are more likely to exit the market in the months following an attack compared to larger outlets.

4.3 Attacks against the press and public interest

This section studies the reaction of the public to an aggression against the press, as this may have an independent effect on publishing. Changes in demand for the content of victimized outlets as a result of an aggression is an important mechanism through which outlet behavior may change. For instance, increased demand for the outlets content would incentivize the press to increase reporting activity.

We test whether public interest in victimized outlets' content changes by regressing *engagement per tweet* on the timing of an attack, in a similar specification as that in model 2. Engagement is defined as the log of likes and re-tweets per tweet (as in Morales, 2019), for an outlet and day combination. Figure 3 shows a 20 percent increase in engagement for the entire sample (and slightly smaller for tweets about violence). Results are similar when considering aleternative specifications (figure A8).

This evidence suggests that the public's engagement with victimized outlets increased, but it is worth highlighting that one potential mechanism that may have induced this increased demand is changes in content. In section 4.4 we show that polarity of content became more negative after the attack, with words that describe the most violent aspects of organized crime being used more often. Nevertheless, these changes in tone are relatively muted, to the point where distinguishing the timing of a tweet based on text content is a difficult task for a modern supervised machine learning algorithm.

Figure 5 depicts coefficient estimates for volume of Google searches by regressing Google Trends volume on event fixed effects and indicator variables for days since murder. We find increases of 10 and 20 percent in search volume for the terms "murder" and "journalist" following an attack, respectively. This is notable since we rely on *national* level searches, while most of the victimized outlets in our sample are small, regional operations, most of them with daily circulation figures below 10,000 units.

Lastly, we do not find any evidence of changes in engagement among journalists located in states where the attacks occurred (Figure A12), even as there are 2% more violence

tweets (figure A11). The public thus appears to care about these homicides, which we interpret as indicating that spikes in engagement among victimized outlets are likely the result of the salience of that event, as opposed to changes in reporting.

4.4 Direct effects of violence on coverage *tone*

An attack on the press may not just affect the volume of news, but also *how* events are covered. These changes may include both the types of news, as well as the language used. We loosely refer to both features as "tone", following standard terminology in the text analysis literature.

To explore these effects we train a modern algorithm, the Multinomial Inverse Regression (MNIR) framework (Taddy, 2013) to identify whether a tweet was published before or after an attack based on natural language. High precision in determining the timing of a tweet is evidence that language changed as a result of an attack. We use our predicted model to fit out-of-sample data and look at "average tone" in the months before and after an attack. We characterize the change in language by identifying the words that contribute the most to determining the timing of a tweet. Finally, we compute the average polarity before and after an aggression. Section 6 formally presents the model.

We consider the tweets published by victimized outlets within 180 days of an attack and estimate the MNIR model separately for all tweets, and tweets with violent content. Because the analysis relies on a medium that follows language conventions more loosely, we perform an aggressive text cleanup. First, we process terms through a so-called "text pipeline" where we drop common prepositions and articles, as these hardly convey useful information. We also lowercase words, eliminate symbols, numbers and hashtags to reduce the feature set to consider. We chose to ignore names of states and municipalities as these terms would be difficult to interpret in the context of the model. Then, we take the root of each word using a standard "stemmer" to account for the fact that the plural and singular version of words are conceptually similar, and to limit the impact of certain types of ortho-

graphic errors. We further avoid including names and words with orthographic errors by only considering stemmed terms from the Spanish version of the Royal Academy for the Spanish Language dictionary (*Real Academia de la Lengua Española*).[14] Lastly, we consider terms that appear in at least 30 tweets to limit over-fitting and prevent some orthographic errors.[15] Surviving features include 2,933 in the entire sample of tweets, and 729 in the violent sample.

We train the MNIR model on this set of words and consider the set of terms with non-zero coefficients, which is approximately one third. Importantly, we ignore approximately 1% of tweets about murdered media workers, as the press covered these events extensively and hence the resulting set of terms picked by MNIR is hardly informative. Specifically, we filter tweets that mention the words "journalist" *and* "murder", or the name of a murdered media worker.

Our first step in the empirical analysis is to assess the ability of the model to correctly predict the timing of a tweet. For this we look at the Sufficient Reduction (SR, denoted by Z_{oi}), which is a projection of the space of counts of words onto the real line. Taddy (2013) describes conditions under which the SR performs as a summary of the available information, pertaining to the dependent variable. Figure A13 presents the distribution of Z for the sample of tweets 180 days before and after an attack. We observe a large fraction of tweets with Z values close to zero for the entire sample (*a*). High values are associated with post-attack, but low values are less indicative of actual timing. This may indicate that new "themes" appeared post-attack. For violence tweets (*b*) the two distributions are more similar with the post-attack distribution slightly shifted to the right. Here, high values of Z are indicative of the post-attack period and vice-versa. Nevertheless, we see a fair amount of overlap in the distributions. Thus, any changes in tone are likely subtle, which could be

[14] To the best of our knowledge there is no machine readable dictionary for the Mexican version of the Royal Academy for the Spanish Language.

[15] Limiting words to those in the Royal Academy for the Spanish Language dictionary does not ensure that no orthographic errors are considered, because terms with an orthographic error might still possess the stem of a correct term.

due to a more homogeneous sample.

We formally test for changes in tone by predicting the probability that a tweet was published post-attack using the SR (Z_{oi}) and a simple linear probability model with calendar month fixed effects and outlet fixed effects. Then, we run an event-study regression of the predicted probability following model 2. Figure 7 reports coefficient estimates. For both samples the DMR model assigns a 0.05% higher probability of post-attack to tweets that were indeed published post-attack, relative to pre-attack. Note that this small coefficient is driven by tweets with SR values close to zero. Ignoring these tweets would lead to much larger estimated effects on tone.

Results are similar when considering calendar month $\times$ state fixed effects.[16] Importantly, the included outlet fixed effects ensure that we capture *within* outlet changes in language, as opposed to any compositional changes that may result from certain outlets reducing their coverage more than others. Lastly, while over-fitting is always a concern in models with a large set of regressors, we find little evidence that this is driving our results: coefficients for $t, \cdots, t+6$, which are estimated using "in-sample" observations, are only slightly larger than subsequent coefficients, which are "out-of-sample". Furthermore, this result is robust to employing 30 days around the day of the event to train the model (figure A14).

To explore how content itself changed we consider the set of words with non-zero coefficients from the MNIR model, which we call $\hat{x}$ (the model produces parsimonious estimates by shrinking term coefficients to zero through a LASSO penalty). We estimate loadings for these surviving features through partial least squares (PLS) by regressing our indicator variable of post-attack, y_{oi}, on the within-tweet fraction of terms with non-zero loading, $\hat{f_{oi}}$. Figure 6 presents the distribution of loadings and frequencies for the full sample of tweets and the sub-sample of violence tweets. The top 30 terms that most *contribute* to identifying the timing of tweet are highlighted, where contribution is defined as the

[16]The specification with calendar months interacted with municipalities is not reported, as there is not enough variation to estimate the model reliably.

product of loading and frequency.

Panel *a* shows that words with large contributions to predicting pre-attack among the full sample have mostly *neutral* meanings such as "news", "state", "north" (probably referencing popular northern music), etc. Perhaps the most loaded word here is "freedom" and "vehicle" which could reference news about drug violence. Words that predict post-attack appear to be more *loaded*: "execution_or_perform", "dead", "president", "victim". Naturally, these differences could stem from changes in the composition of news: if coverage of violence increases proportionally after an attack we would expect the MNIR framework to pick words with violent loading as post-attack predictors.

Panel *b* limits this issue by presenting results for the sample of tweets with violent content. Terms that predict pre-attack are clearly more loaded now with surviving features such as: "alleged", "crime", "war", "army" and "agency". Thus, it appears that outlets report more about crime and operations by the military and law enforcement prior to an attack. Afterward, attention shifts to the most visible signs of cartel violence: "confront_or_in_front", "forced_disappearance", "body", "victim", "execution_or_perform".

We further test this intuition by studying changes in polarity as a result of an aggression. We rely on a polarity dictionary from Brooke, Tofiloski and Taboada (2009). This dictionary considers adjectives, nouns, adverbs and verbs from reviews for hotels, movies, music, phones, washing machines, books, cars, and computers from the website Ciao.es. Semantic orientation values were assigned by a Spanish native speaker, and compared to crowd-source classification via Mechanical Turk. Other methods tested by the authors such as Vector Machine Learning (VME) and automatic translation of an existing English dictionary performed worse than this baseline dictionary, which has an accuracy of 74.50%. Polarity is codified as an integer between -5 and 5, inclusive.

Figure 8 shows that polarity became more negative after an attack. For the full sample we observe -.05 to -.1 more negative loadings. Among violence tweets we see initially more positive polarity one to two months after an event, followed by more negative average

polarity. Interestingly, these patterns are starker when controlling for calendar month $\times$ state fixed effects, which implies that targeted outlets experienced a differential decrease in polarity compared to outlets in the same state. Figure A15 shows the distribution of terms by polarity before and after an aggression. In panel *a* 10% of all classified terms in the full sample of tweets had a polarity of 5 (the maximum possible) before the event, as well as a similar number with polarity 3 and 4. Approximately 8% of terms had a polarity of -4, the lowest observed. Following an attack, 20% of tweets used very negative language (-5 polarity), while no tweet used very positive language (5). Among tweets with violent content (panel *b*) we observe more neutral language before the attack, with polarity ranging from -4 to 3. Afterwards, we observe 12% of terms exhibiting very negative polarity (-5), but also 4% with very positive polarity (5). This likely explains the initial muted change among violence tweets.

4.5 Indirect effects

In this section we examine indirect effects of the attack on non-victimized journalists. In particular, we examine whether there are localized spillover effects by examining a narrow set in both physical and social distance from the killing: looking at journalists in the same state and neighboring in their Twitter network.[17] In the presence of indirect behavioral responses we may observe either an increase in tweets denouncing the aggression, and perhaps criticizing the government's response (backlash), or a reduction in publishing activity out of fear of becoming victimized (indirect deterrence).

Like-minded individuals are both more likely to follow each other on Twitter (Halberstam and Knight, 2016) and to cover similar content. As such, individuals who followed the victimized outlet on Twitter may respond (differentially) to the killing relative to other journalists. In figure 4 we compare the behavioral response of journalists in the victim's network *and* state, to other journalists. The model we estimate is given by:

[17] Figure A18 shows the Twitter network for the accounts in our dataset.

$$
\begin{aligned}
y_{jsft} &= \sum_{k=-6}^{12} \beta_k \times periodsSinceKilling_k \times inState_s \times inFriends_f \\
&+ \lambda_j + \lambda_{sm} + \lambda_{fm} + \epsilon_{jsft}
\end{aligned} \tag{3}
$$

where y_{jsft} is the log tweets by journalist j, for time periods t, in state s, who follows victimized outlet f on Twitter; the rest of the notation is the same. This triple-difference model includes both state $\times$ month fixed effects (λ_{sm}) and victimized-outlet-follower $\times$ month fixed effects (λ_{fm}). The coefficients of interest capture localized spillovers for followers of the victimized outlet who are located in the same state, relative to other journalists in the state who did not follow the victimized outlet, and to other followers of the victimized outlet but who are located in different states. We observe a sustained reduction in twitter activity starting in the second month after the event (Panel *a*) of around 10 percent in the short to medium run, and larger in the long run. Though the estimates are noisier, there appears to also be a reduction in coverage of violence (Panel *b*) of around 5 percent. The results suggest that the killings are not only effective in reducing coverage of the victimized outlets, but also of their nearby peers.

4.6 Long-run effects

The previous sections showed that targeted outlets scaled down coverage in the short to medium run, while spillover effects were modest. Attacks against the press may however affect press activity significantly in the long run as outlets are able to allocate fewer resources or exit the market altogether. This section tests whether states with more aggression saw comparative decreases in the size of the press. We rely for this on the Mexican Census that started recording individuals working as journalists in 2010. We show here that more dangerous states experienced comparative reductions in the number of journalists.

This section also presents evidence of changes in demographic and labor market characteristics of these individuals. As a comparison group, we include workers in the sample whose occupational codes in the census are close to that of journalists, such as accountants, researchers, psychologists, artists and performers. The sample of analysis is restricted to these occupational categories.

Difference-in-differences: The share of journalists

We first implement a difference-in-differences specification to examine whether higher violence against the press is associated with changes in the share of journalists. In particular, we rely on regressions of the following form

$$y_{ist} = \alpha + \beta \times violence_s \times post_t + \gamma_t + \gamma_s + \varepsilon_{ist} \tag{4}$$

where the outcome of interest y_{ist} is an indicator equal to one if individual i, in state s, for census t, reports being a journalist as their occupation. The treatment variable $violence_s$ includes the number of journalists killed in state s between 2010 and 2015 (or alternatively, an indicator equal to one if at least one journalist was killed in the state), and the $post_t$ indicator equals one if the observation corresponds to the 2015 sample. Our preferred specification includes state and year fixed effects γ.

Results are reported in table 1. We observe that the share of journalists decreased differentially between 2010-2015 in more violent states. Since about 3.5 percent of workers in the selected 2010 sample worked for the press, this implies that in states with any killing, the share of journalists decreased by 25 percent ($\beta = -0.0087$, column 4), relative to states where no killings took place (and to the comparison occupations). Results are similar with or without state and year fixed effects, and when we consider only wage earners. Long-run effects are thus in line with our findings for the short to medium run.

Triple-differences: Characteristics of Mexican journalists

We provide further evidence of changes in the operation of the press in the country by studying whether the pool of individuals who decided to work as journalists in violent states changes as well. We rely for this on triple-differences regression models

$$\begin{aligned} y_{isto} = \alpha &+ \beta_0 \times violence_s \times post_t + \beta_1 \times violence_s \times journalist_o \\ &+ \beta_2 \times journalist_o \times post_t + \beta_3 \times violence_s \times journalist_o \times post_t \\ &+ \gamma_o + \gamma_t + \gamma_s + \varepsilon_{ist} \end{aligned} \tag{5}$$

where the outcomes of interest y_{isto} include demographic and labor market characteristics, such as number of kids, marriage status, years of education, age and income, among others, for individual i, in state s, for census t, and occupation o. The coefficient of interest, β_3, measures changes in the outcome of interest for journalists in violent states in 2015, relative to the comparison group.

Results are reported in table 2. In states with higher number of media workers killed, journalists in 2015 are less likely to be married, have children, live in urban areas, and earn less money. The coefficients also suggest they are on average less educated and younger, though these are imprecisely measured and not statistically significant. One possible interpretation of these findings is that individuals with these characteristics are more willing to engage in this dangerous profession.

4.7 Coverage by the national press

In this section we explore how the national press covered the homicides of its colleagues. We document that these events received plenty of attention, but also that the press did not generally attribute these attacks to any criminal organization, either due to lack of knowledge or for fear of reprisal. We also show that the aggression did not lead to permanent

changes in reporting, which could be expected as large national outlets are not generally targeted by criminal organizations.

Figure A19 panel *a* shows an event study of mentions of municipalities by outlets in the EFIC database before and after the homicide of a journalist. We considered the set of municipalities with non-ambiguous names[18], which is approximately 90% of them (municipalities are often named in Náhuatl and other indigenous languages, and thus we are unlikely to miss-classify a Spanish word in a news item). We matched in this fashion newspaper articles, radio segments and TV programming to municipalities in the country. We find that following an attack there's a .17 log point increase in mentions of the municipality. No effect subsidies after the first month, and we find no evidence of pre-trends before the event.

In a similar fashion we look at mentions of pairs of municipalities and keywords, such as "hitman", "Sinaloa Cártel" and "Jalisco Cártel".[19] Panel *b* shows an increase in mentions of the term *hitman*, but we do not find any change in the number of references to criminal organizations (panel *c* and *d*).

Figure A20 depicts the distribution of the SR statistic, where we train the MNIR model to identify the timing of a tweet from natural language from EFIC news articles that were published 30 days before and after an attack took place in a given *state*. Panel *a* shows some post-attack news items have high values of SR, but some have very low values as well. There's also a high degree of overlap in the two distributions around zero, which means that there's little informational content in the language of these news items to distinguish their timing. In panel *b* we show an event study of the predicted probability of post-attack based on the SR. There's an approximate 0.4% jump in probability after an attack, but this effect quickly subsidies (for ease of interpretation we normalized the coefficients with re-

[18]Some municipalities share names with one or more municipality, or are commonly used nouns.

[19]The last two are currently the largest criminal organizations in the country, and are considered by the US government as the main criminal threat faced by that country. We ran similar regressions for other known criminal groups and found no increase in mentions either. Results are available from the authors upon request.

spect to -2). The fact that coefficients from 2 through 6 are all close to zero suggest that either there are no persistent effects in tone, or (more likely) that the jump in probability that we measured was driven by over-fitting in the MNIR model, and thus it's possible that tone did not change at all.

5 Conclusion

With the start of the Mexican War on Drugs in 2006 the annual number of journalists killed in the country doubled, while the total number of murders increased almost three-fold. This article studies how the news media in Mexico changed its coverage of violence in response to high levels of victimization suffered by journalists and other media workers.

We document that a large majority of victims covered crime. While 97% of these murders remain unsolved, press reports suggest that drug-trafficking organizations planned and carried out these homicides. Victims were affiliated to small, local news outlets that reported on local crime with a level of detail that larger national outlets do not generally provide.

Following an attack these outlets reduced their coverage sharply. This occurred even as public interest in their content peaked, which underscores how effective violence was at censoring. Smaller outlets, that loose a higher share of their employees to an attack, were especially at risk, with a full 20% of them exiting the market for crime news permanently within a month of the homicide.

In the long run, these homicides reduced the supply of news. States that reported the murder of a media worker saw reductions in the number of active journalists. Individuals who remained in the profession were less likely to be married, have kids, less likely to live in urban areas and had lower income, which could be indicative of lower risk aversion.

Our measured effects are not only the result of mechanical effects, but also behavioral responses to a more dangerous environment. First, the tone and language of coverage changed permanently. Post-attack tweets from targeted outlets underscored the most vi-

olent aspects of organized crime, such as extra-judicial executions and confrontations. In doing so, polarity tended towards more negative. Second, journalists in the same network and state as the victim (thus likely more at risk) reduced their Twitter activity comparatively.

This article contributes to the literature by documenting how effective violence is at censoring the press in the context of Mexico and –in doing so–, reports on yet another externality of drug trafficking. Limiting information, particularly local information that may not be reported elsewhere, hurts the well functioning of a democracy and may reduce incentives of public officials to combat organized crime.

References

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. 2015. "Radio and the rise of the Nazis in Prewar Germany." *The Quarterly Journal of Economics*, 130(4): 1885–1939.

Beattie, Graham, Ruben Durante, and Brian Knight. 2017. "Advertising Spending and Media Bias: Evidence from News Coverage of Car Safety Recalls." *Working Paper*.

Brooke, Julian, Milan Tofiloski, and Maite Taboada. 2009. "Cross-linguistic sentiment analysis: From English to Spanish." 50–54.

Chiang, Chun Fang, and Brian Knight. 2011. "Media bias and influence: Evidence from newspaper endorsements." *Review of Economic Studies*, 78(3): 795–820.

Dellavigna, Stefano, and Ethan Kaplan. 2007. "the Fox News Effect: Media Bias and Voting." *The Quarterly Journal of Economics*, 122(3): 1187–1234.

di Tella, Rafael, and Ignacio Franceschelli. 2011. "Government advertising and media coverage of corruption scandals." *American Economic Journal: Applied Economics*, 3(4): 119–151.

Durante, Ruben, and Brian Knight. 2012. "Partisan control, media bias, and viewer responses: Evidence from Berlusconi's Italy." *Journal of the European Economic Association*, 10(3): 451–481.

Gentzkow, Matthew, and Jesse Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica*, 78(1): 35–71.

Halberstam, Yosh, and Brian Knight. 2016. "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter." *Journal of Public Economics*, 143: 73–88.

Holland, Bradley E., and Viridiana Rios. 2017. "Informally Governing Information: How Criminal Rivalry Leads to Violence against the Press in Mexico." *Journal of Conflict Resolution*, 61(5): 1095–1119.

Hughes, Sallie, and Mireya Márquez-Ramírez. 2018. "Local-level authoritarianism, democratic normative aspirations, and antipress harassment: Predictors of threats to journalists in Mexico." *The International Journal of Press/Politics*, 23(4): 539–560.

Knight, Brian, and Ana Tribin. 2019. "The Limits of Propaganda: Evidence from Chavez's Venezuela." *Journal of the European Economic Association*, 17(2): 567–605.

Morales, Juan S. 2019. "Legislating during war: Conflict and politics in Colombia." *Working Paper*.

Morales, Juan S. 2020. "Perceived Popularity and Online Political Dissent: Evidence from Twitter in Venezuela." *The International Journal of Press/Politics*, 25(1): 5–27.

Pan, Jennifer, and Alexandra A. Siegel. 2019. "How Saudi Crackdowns Fail to Silence Online Dissent." *American Political Science Review*, 109–125.

Ramírez Alvarez, Aurora. 2017. "Media and Crime Perceptions: Evidence from Mexico."

Relly, Jeannine E, and Celeste González de Bustamante. 2014. "Silencing Mexico: A study of influences on journalists in the Northern states." *The International Journal of Press/Politics*, 19(1): 108–131.

Reuter, J., and E. Zitzewitz. 2006. "Do Ads Influence Editors? Advertising and Bias in the Financial Media." *The Quarterly Journal of Economics*, 121(1): 197–227.

Salazar, Grisel. 2019. "Strategic allies and the survival of critical media under repressive conditions: An empirical analysis of local Mexican press." *The International Journal of Press/Politics*, 24(3): 341–362.

Snyder, James M., and David Stroemberg. 2010. "Press Coverage and Political Accountability." *Journal of Political Economy*, 118(2): 355–408.

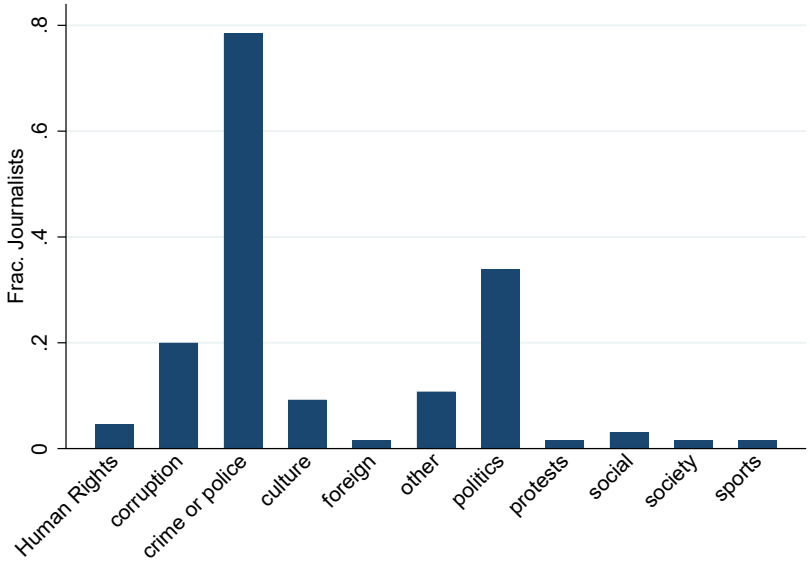
Stanig, Piero. 2015. "Regulation of speech and media coverage of corruption: An empirical analysis of the Mexican Press." *American Journal of Political Science*, 59(1): 175–193.

Taddy, Matt. 2013. "Multinomial inverse regression for text analysis." *Journal of the American Statistical Association*, 108(503): 755–770.

Valdez, Javier. 2016. *Narco periodismo: la prensa en medio del crimen y la denuncia.* Aguilar.

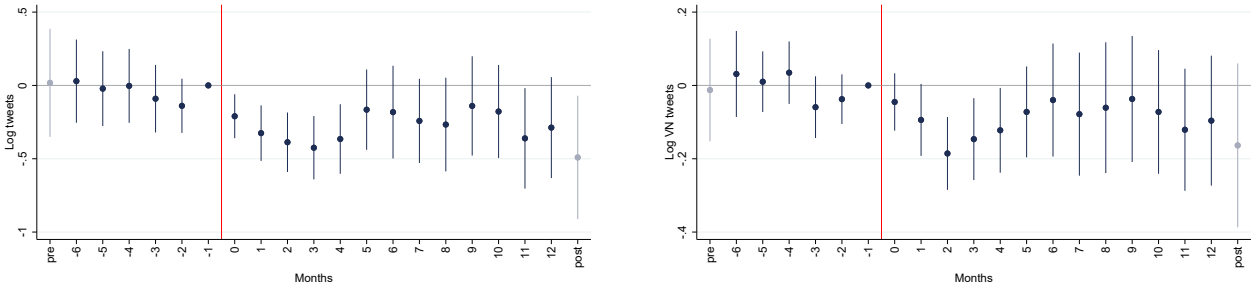
6 Tables and figures

Figure 1: Subjects covered by victims



Note: own construction based on reports from CPJ as of October 15, 2017.

Figure 2: Direct effects of an attack on volume of coverage of victimized outlet

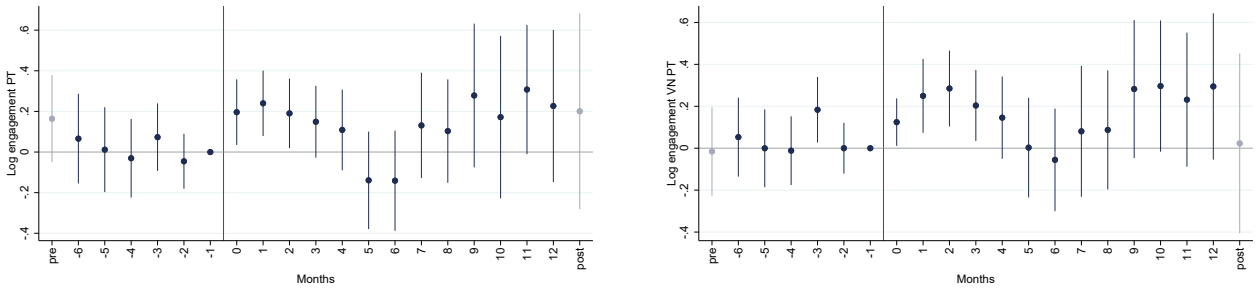


(a) All tweets, month × state FE

(b) Violence tweets, month × state FE

Note: All regressions include outlet fixed effects, and states where outlets are located × calendar months fixed effects. Robust standard errors clustered by outlet.

Figure 3: Direct effects of an attack on Twitter engagement for victimized outlets

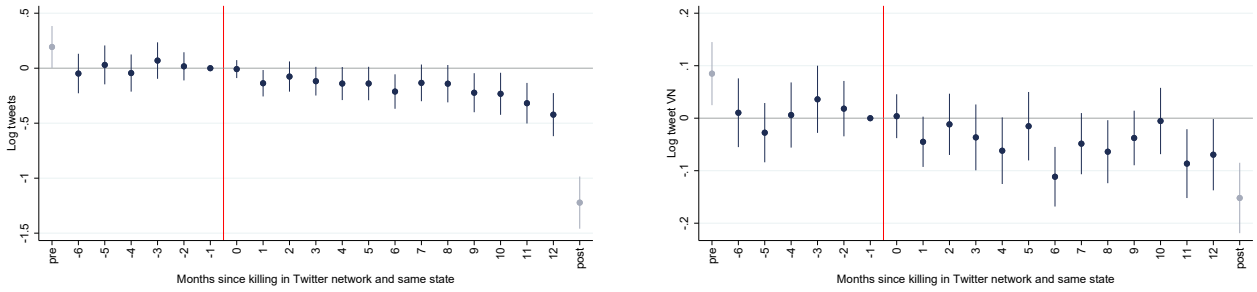


(a) All tweets, month × state FE

(b) Violence tweets, month × state FE

Note: engagement is defined as likes and re-tweets normalized by the number of tweets for a given outlet-day combination. All regressions include outlet fixed effects, and states where outlets are located × calendar months fixed effects. Robust standard errors clustered by outlet.

Figure 4: Indirect effects on volume of coverage for journalists following victimized outlets on Twitter and located in the same state

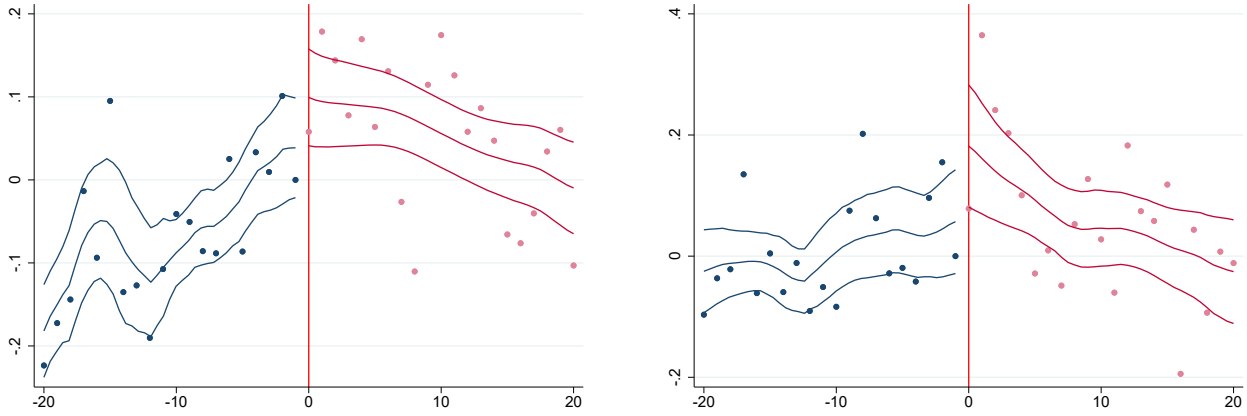


(a) All tweets

(b) Violence tweets

Note: Regressions compare Twitter activity of journalist in the same state that followed on Twitter murdered media workers, and those that did not.

Figure 5: Google trends search volume

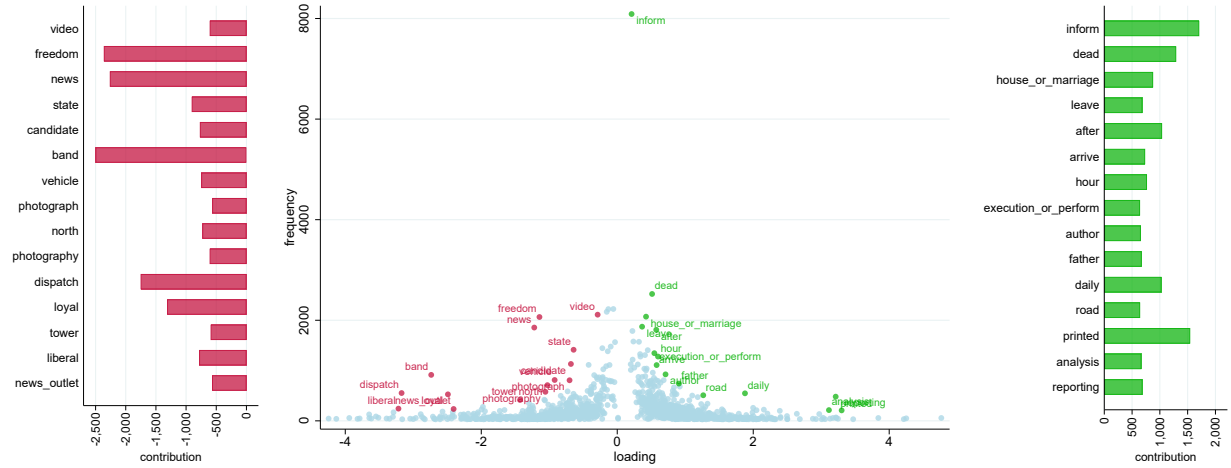


(a) Keyword: *murder*

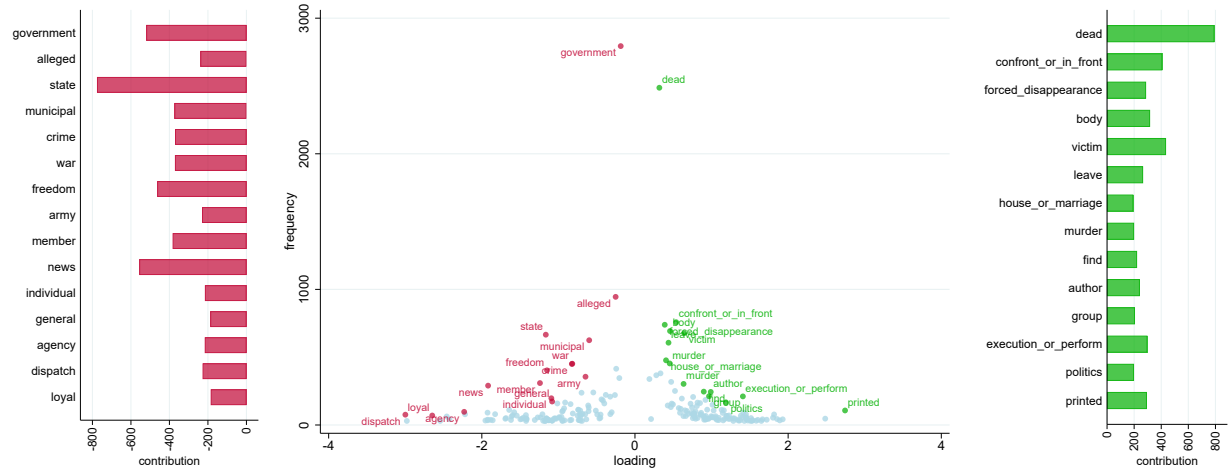
(b) Keyword: *journalist*

Note: daily national Google search volume for "murder" and "journalist" 20 days before and after the murder of a journalist. Includes event fixed effects. Epanechnikov kernel plot with bandwidth based on a rule-of-thumb.

Figure 6: Terms that most predict timing of a tweet



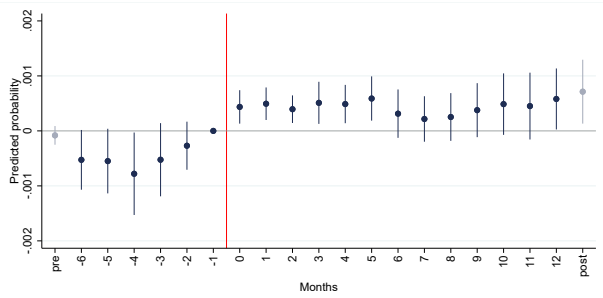
(a) All tweets



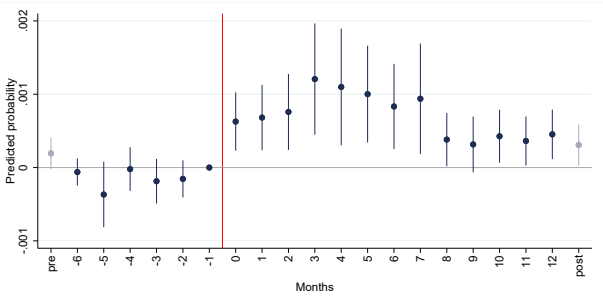
(b) Violence tweets

Note: We train a MNIR model on tweets 180 days before and after an attack to predict whether a tweet was published post-attack. These plots show the distribution of loadings and frequencies of the feature set of words with non-zero loading from the MNIR model. Highlighted are the top 30 terms that better predict timing based on their total contribution, defined as the product of loading and frequency. Tweets referrencing murders of journalists were omitted.

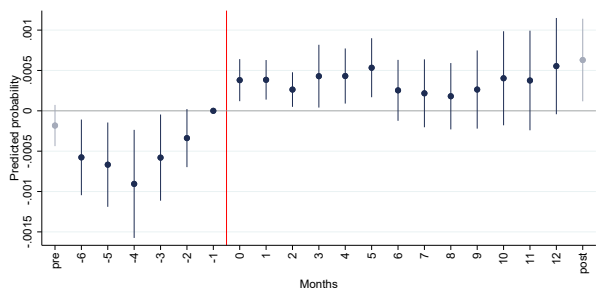
Figure 7: Tone of coverage among targeted outlets



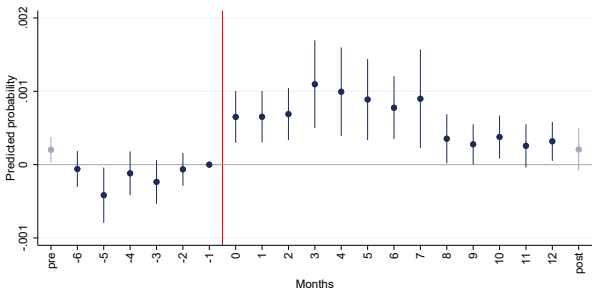
(a) All tweets, calendar month FE



(b) Violence tweets, calendar month FE



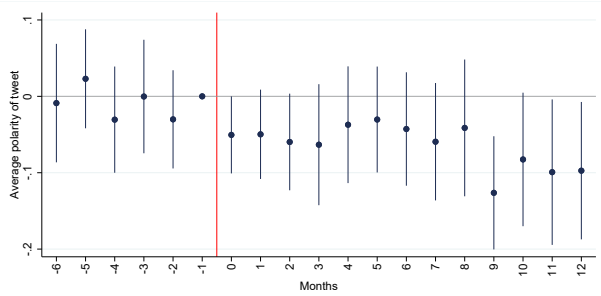
(c) All tweets, calendar month $\times$ state FE



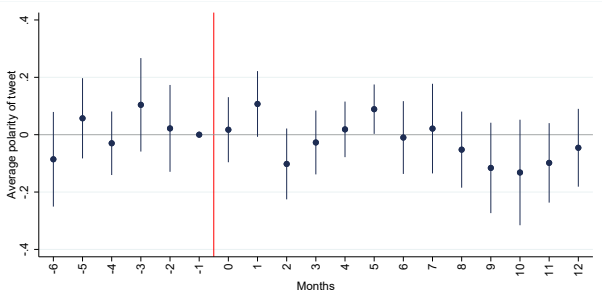
(d) Violence tweets, calendar month $\times$ state FE

Note: we estimate a logit model where we regress the post-attack indicator variable y_{oi} on Z_{oi}, the *Sufficient Reduction* (SR) of the count space of words. This process is referred to as *forward regression* by Taddy (2013). These figures depict event study estimates of the average monthly predicted probabilities of post-attack. Lower values indicate that text content was more similar to the pre-attack content, while higher values suggest that content was more similar to the post-attack content. Both the MNIR model that generates the SR and the logit model are estimated using 180 days worth of tweets before and after a homicide, thus coefficient estimates for $7, \cdots, 12$, as well as *pre*, *post* are out-of-sample. Robust standard errors clustered by outlet.

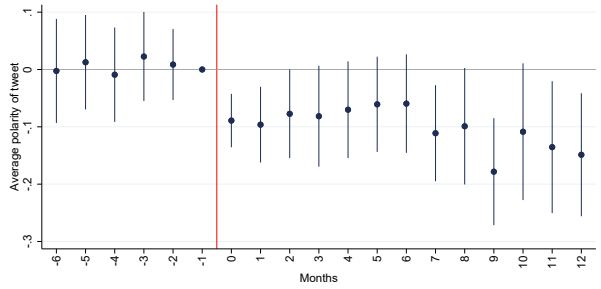
Figure 8: Changes in polarity following an attack



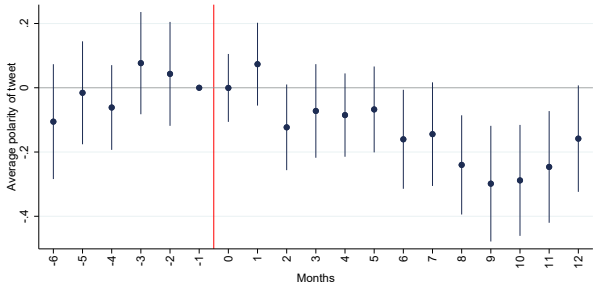
(a) All tweets, calendar month FE



(b) Violence tweets, calendar month FE



(c) All tweets, calendar month × state FE



(d) Violence tweets, calendar month × state FE

Note: we consider the set of terms with non-zero loading from the MNIR estimation, that are also considered in Brooke et al. (2009) polarity dictionary. Polarity is computed as the arithmetic average of polarity of tweets. We control for outlet fixed effects. Panel *a* and *b* control for calendar month fixed effects, while *c* and *d* interact these with state indicators where the outlet is located. Robust standard errors clustered by outlet.

Table 1: Relationship between violence against journalists and share of journalists

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
post x Nr. MW murd.	-0.0010*	-0.0009*			-0.0011*	-0.0010**		
	(0.0005)	(0.0004)			(0.0005)	(0.0005)		
post x log hom.	0.0001	-0.0001	0.0003	0.0000	-0.0001	-0.0005	0.0001	-0.0003
	(0.0014)	(0.0014)	(0.0016)	(0.0016)	(0.0014)	(0.0014)	(0.0016)	(0.0016)
post x Any. MW murd.			-0.0092***	-0.0087***			-0.0107***	-0.0100***
			(0.0032)	(0.0030)			(0.0036)	(0.0034)
N	117673	117673	117673	117673	101421	101421	101421	101421
N-clusters	32	32	32	32	32	32	32	32
State FE	no	yes	no	yes	no	yes	no	yes
Year FE	no	yes	no	yes	no	yes	no	yes
Only wage earners	no	no	no	no	yes	yes	yes	yes

Notes: Outcome is an indicator equal to one if individual reports being a journalist. Robust standard errors clustered at the state level in parenthesis. Significance levels shown below *p<0.10, ** p<0.05, ***p<0.01.

Table 2: Relationship between violence against journalists and census outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	School yrs	Married	Kids	Age	Urban	Male	Income
post x Nr. MW killed x journ.	-0.0355	-0.0104**	-0.0106**	-0.0191	-0.0026**	0.0035	-0.0215***
	(0.0342)	(0.0047)	(0.0040)	(0.0798)	(0.0010)	(0.0038)	(0.0071)
post x log hom. x journ.	-0.1123	-0.0188	-0.0277	0.3898	-0.0011	-0.0231	0.0349
	(0.1123)	(0.0328)	(0.0336)	(0.4736)	(0.0088)	(0.0247)	(0.0569)
N	117673	117673	117673	117673	117673	117673	101421
N-clusters	32	32	32	32	32	32	32
State FE	yes	yes	yes	yes	yes	yes	yes
Year FE	yes	yes	yes	yes	yes	yes	yes

Notes: Robust standard errors clustered at the state level in parenthesis. Significance levels shown below *p<0.10, ** p<0.05, ***p<0.01.

Appendix

Data censoring

As mentioned in section 3.3, we observe a large number of outlet-day combinations with 20 tweets, which might indicate censoring in the data. We address this by estimating an event-study specification considering the number of combinations of outlets-days with data-censoring as the left-hand side variable (figure A6). We measure a statistically insignificant 30 percent reduction in the specification with month fixed effects, and a negligible and insignificant increase in the baseline specification with month and state fixed effects. Any bias introduced by data-censoring thus likely underestimates the true reduction in Twitter activity.

To further confirm these patterns, we look at the mean timing of tweets, defined as the log of the average number of seconds elapsed until the end of the day for tweets within a given day (larger values indicate that an outlet published tweets earlier in the day) by a given outlet. As long as the distribution of tweets during the day remains unaffected by an attack, lower values (i.e., tweets published closer to the end of the day) indicate a higher probability that a particular day had more tweets, and vice-versa (Morales, 2020). Figure A7 shows a similar pattern to our main specification (figure 2): an increase in average time elapsed (i.e., fewer tweets) following the attack that peaks three months later.

Online newspaper articles

Out of the 104 victimized outlets that we consider we found a URL for 86 of them. We used the Google Custom Search Engine (CSE) API to find articles for those outlets using the following list of keywords: *narco*, *ejecución* (execution), *fosas* (illegal grave), *cartel*. Our queries produce a similar set of results as one would obtain from querying: **site:outlet-webpage.com "keyword" range:**$date_1$**-**$date_2$. Note that Google CSE is context-aware, such

that it will return matching articles even if they do not include the specific keywords in the article, provided Google's proprietary algorithms determine that it is relevant to the query. Google CSE produces a maximum of 100 results for a given query. This limitation is problematic because we would underestimate the number of matches for queries with more than 100 hits. To address this issue we restrict our queries to 30-day windows. Out of the 10,740 outlet $\times$ 30-day window $\times$ term queries, we only encountered one with 100 hits. Another empirical issue that we faced was that some newspaper website's were no longer online. In those cases Google CSE would not return any matching article URLs. This was true for 18 out of 86 newspapers in the sample. In total we found the URLs for 98,595 articles. For a majority of newspapers we found less than 2,000 articles, whereas for some in the right tail of the distribution we found more than 7,000.

Unfortunately, the CSE database has significant drawbacks. First, the heuristics that we use to find the date of articles work for only a small subset of them, either because the heuristic failed or, more commonly, because the article does not contain a date. We thus rely on the 30-day query itself to assign dates, which reduces precision (for articles where we can retrieve a date Google's date range is accurate in more than 95% of cases). Two outlets are included both in this database, and in our national outlet database (**EFIC**). We show in section 3.2 that EFIC strongly correlates with national homicides in the country, which allows us to test the quality of the CSE data by comparing them with EFIC. Unfortunately, CSE correlates weakly with the EFIC database, which indicates issues with CSE. To limit them we consider as dependent variable in our event-study estimates an indicator for outlet $\times$ 30-day periods with *any* articles.

Figure A9 depicts our estimates. We observe a fall of approximately .2 log points in coverage after an attack that peaks 4 months after the aggression. The coefficient for $t = 1$ is not statistically different than $t = -1$, but this could be due to initial reporting about the attack itself. Our estimates are thus in line with those using Twitter in section 4.2.

The DMR model

Following Taddy (2015), define c_{oi} to be the vector of counts of words across d possible categories, indexed by j, for observation i of outlet o, which sum up to $m_{oi} = \sum_j c_{oij}$. Let y_{oi} be a scalar equal to one if observation oi corresponds to the post-attack period of outlet o, and zero otherwise. The model we try to estimate assumes that every document count c_{oij} has been drawn independently Poisson with intensity $e^{\eta_{oij}}$, $P_o(e^{\eta_{oij}})$. The joint document likelihood for c_{oi} factorizes as

$$p(c_{oi}) = \Pi_j P_o(c_{oij}; e^{\eta_{oij}}) = MN(c_{oi}; q_{oi}, m_{oi}) P_o(m_{oi}; \Lambda_{oi}) \tag{6}$$

where $\eta_{oij} = \alpha_j + \mu_{oi} + y_{oi}\psi_j$, $\Lambda_{oi} = \sum_{k=1}^{d} e^{\eta_{oik}}$, and q_{oij} is the probability of category j for observation oi. This model has a negative log likelihood that is proportional to

$$\sum_{j=1}^{d} \sum_{i=1}^{n} \left[e^{\mu_i + \eta_{oij}} - c_{oij}(\mu_{oi} + \eta_{oij}) \right] \tag{7}$$

with a gradient on μ_{oi} of $g(\mu_{oi}) = e^{\mu_{oi}} \sum_j e^{\eta_{oij}} - m_{oi}$. In practice, we use the plug-in estimator $\hat{\mu}_{oi} = \ln m_{oi}$. Taddy (2015) shows that this strategy performs well empirically. Under those assumptions each separate Poisson regression has negative log likelihood proportional to

$$l(\alpha_j, \phi_j) = \sum_{i=1}^{n} \left[m_i e^{\alpha_j + y_i \phi_j} - c_{oij}(\alpha_j + y_i \phi_j) \right] \tag{8}$$

we are abstracting here from the outlet to which each observation belongs to, and thus we are "stacking" all observations across all outlets, with $n = \sum_o n_o$ where n_o denotes the number of observations for outlet o. We estimate α_j, ϕ_j through "Gamma-Lasso"

$$\hat{\alpha}_j, \hat{\phi}_j = \arg\min_{\alpha_j, \phi_j} \left\{ l(\alpha_j, \phi_j) + n\lambda \sum_{k=1}^{p} \omega_{jk} |\phi_{jk}| \right\}, \text{ where } \lambda, \omega_{jk} \geq 0 \tag{9}$$

we chose $\omega_{jk} = 1$ as in standard Lasso, as this leads to a less thinner set of surviving words. λ is selected according to *corrected* Akaike Criterion

$$-2l(\hat{\alpha}_j, \phi_j) + 2df_j \frac{n}{n - df_j - 1} \tag{10}$$

where df_j is the estimated degrees of freedom to fit $\{\hat{\alpha}, \hat{\phi}_j\}$. Once we estimate this model we perform a "sufficient reduction" (SR) where we project the high-dimensional space of words into the real line.

$$z_{oi} = \Phi c_{oi} \tag{11}$$

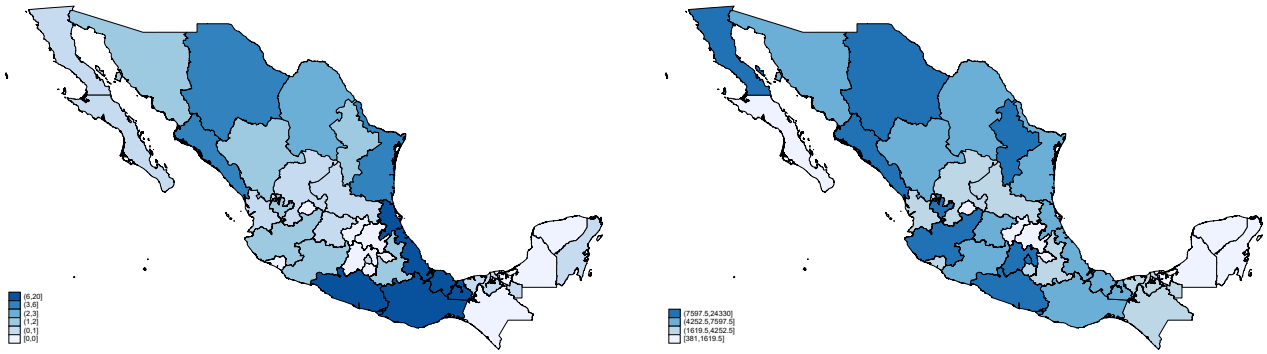
Taddy (2015, 2013) shows how this SR captures all relevant information on y_{oi} from c_{oi} in the sense that c_{oi} does not add any new information on y_{oi}, *controlling for* z_{oi}.

Appendix tables and figures

Table A1: Summary statistics for journalists in the Mexican census

	Mean (no killings)	SD	Mean (>0 killings)	SD	diff p-value
Census year	2012.526	2.5	2012.581	2.5	.503
Male	.554	.497	.685	.465	0
Age	39.478	13.179	39.147	13.883	.461
Yrs school	14.959	2.823	14.192	3.339	0
Married	.468	.499	.571	.495	0
Has children	.337	.473	.391	.488	.001
Christian	.401	.49	.434	.496	.046
Moved state (pr. 5 yrs)	.098	.297	.066	.248	0
Urban	.936	.245	.878	.327	0
Wage worker	.717	.45	.762	.426	.002
No income reported	.14	.347	.131	.337	.441
Log income	9.05	.813	8.753	.81	0
N	2379	.	1505	.	.

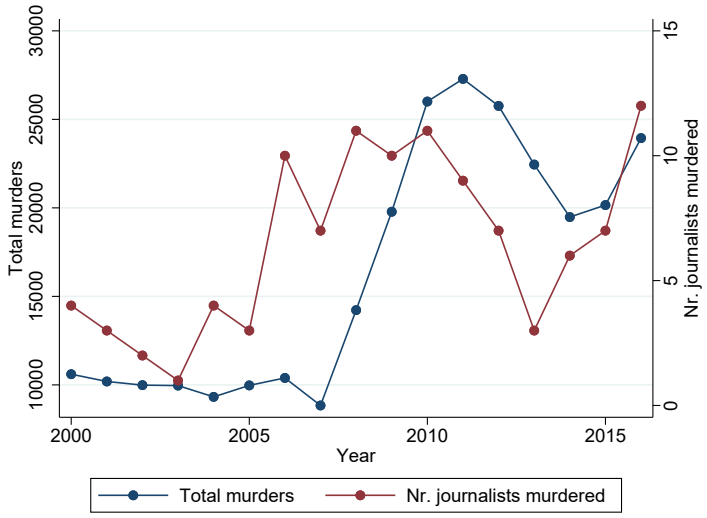
Figure A1: Homicides in the country (2009-2017)



(a) Journalists (b) All homicides

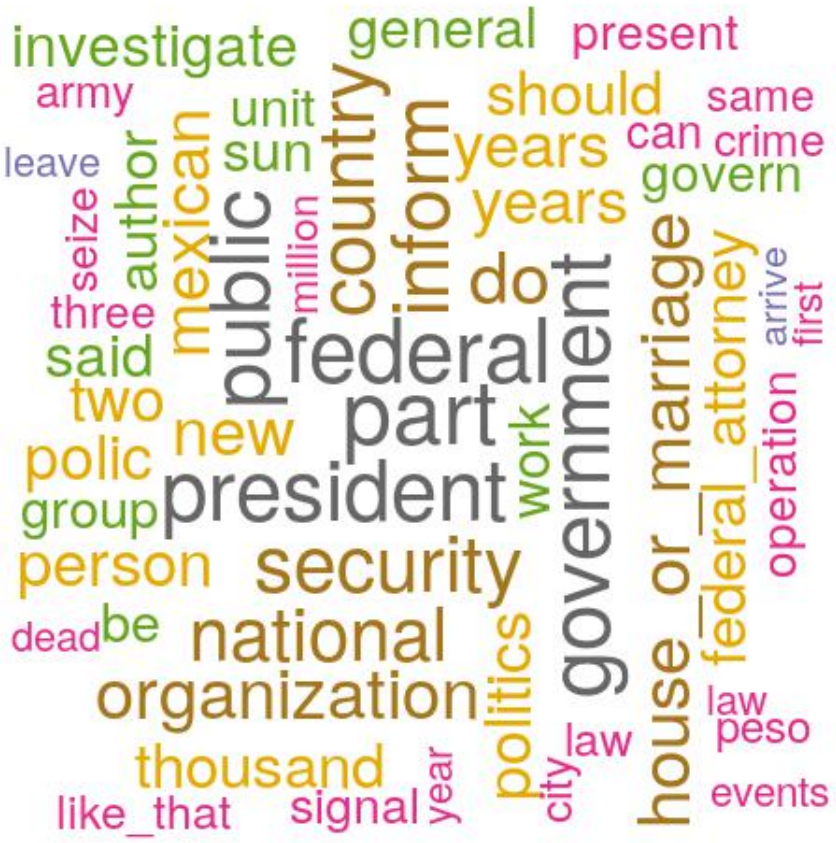
Note: panel *a* depicts homicides of journalists and panel *b* total homicides in the 32 states of Mexico between 2009 and 2017.

Figure A2: Homicides



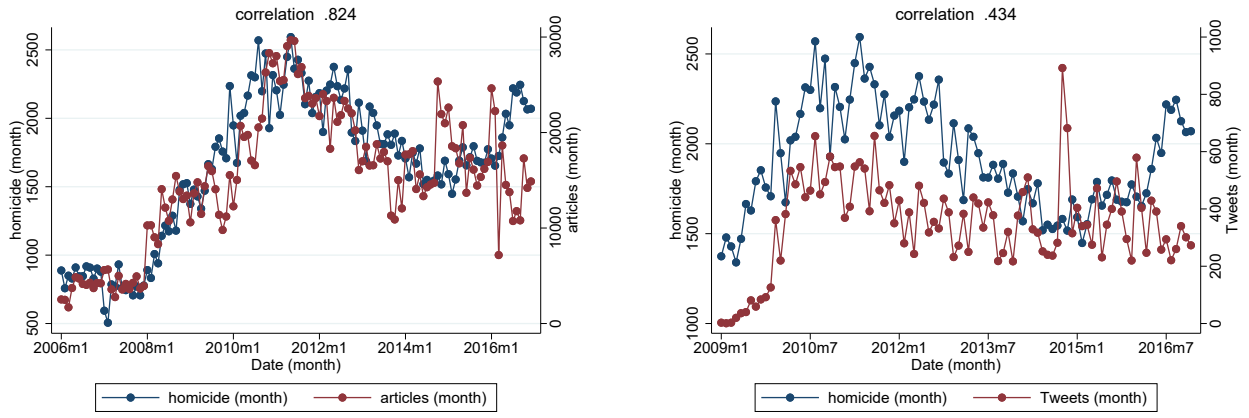
Note: This figure presents annual homicides in the country among the general population and the press.

Figure A3: Most common words (EFIC)



Note: This plot depicts the relative frequency of the 60 most common terms in EFIC. Larger size indicates higher frequency. Terms in the same color have similar frequencies.

Figure A4: Monthly articles/tweets and homicides in the country

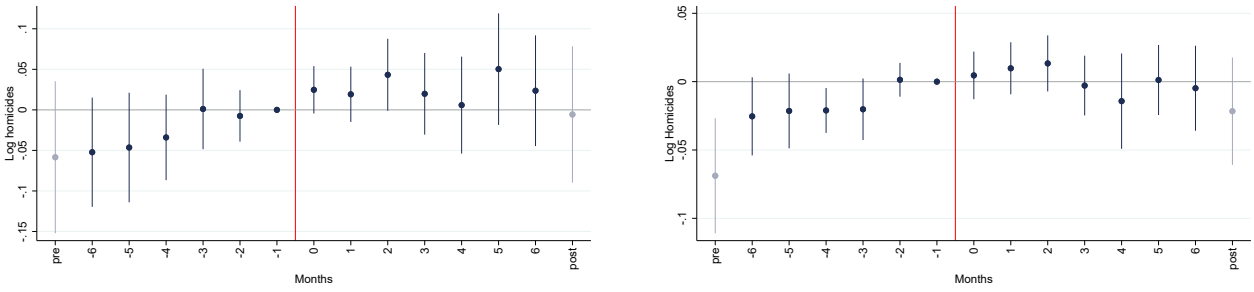


(a) EFIC

(b) Tweets about violence

Note: This figure depicts the time series of monthly number of articles from the "national press" (EFIC, panel *a*) and the number of tweets by the most important journalists in the country (panel *b*) against the number of total homicides. Panel *b* likely underestimates the true correlation, due to censoring (see section 3.3).

Figure A5: Homicides around murders of media workers

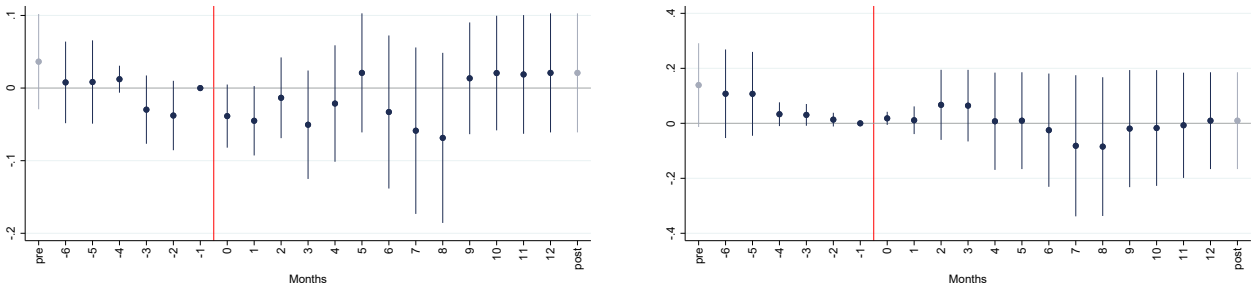


(a) State

(b) Municipality

Note: Regressions include month fixed effects, and state or municipality fixed effects, respectively. Homicide figures exclude murders of media workers. Robust standard errors clustered by state or municipality.

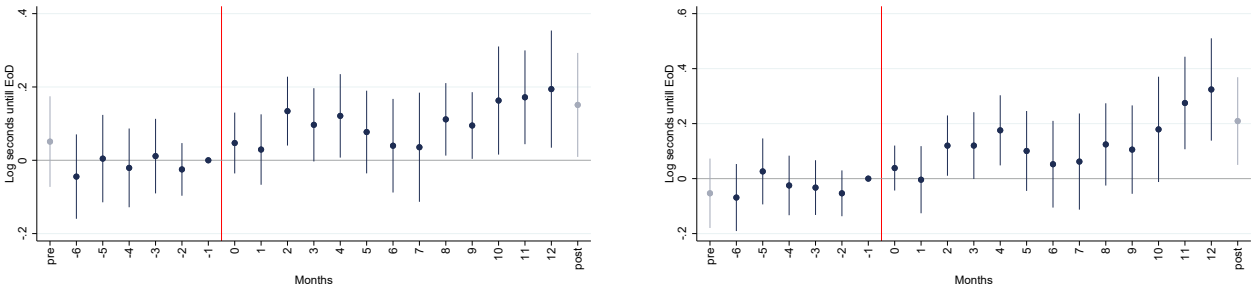
Figure A6: Victimization and censoring among targeted outlets



(a) Calendar month FE (b) Calendar month × state FE

Note: Sample considered are outlets × days with 20 tweets, as this might indicate censoring. All specifications include outlet fixed effects. Robust standard errors clustered by outlet.

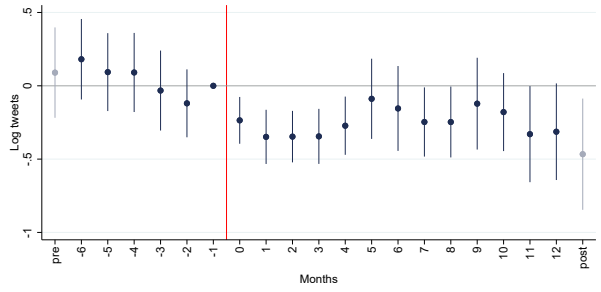
Figure A7: Average timing of tweets of victimized outlets



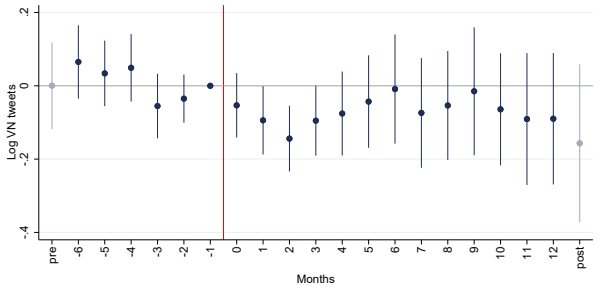
(a) Calendar month FE (b) Calendar month × state FE

Note: Dependant variable is the log of the average of seconds elapsed for each observed tweet since the start of day. Higher values indicate that tweets were published earlier in the day. Under the assumption that distribution is independent of the attack, earlier values indicate fewer tweets. All specifications include outlet fixed effects. Robust standard errors clustered by outlet.

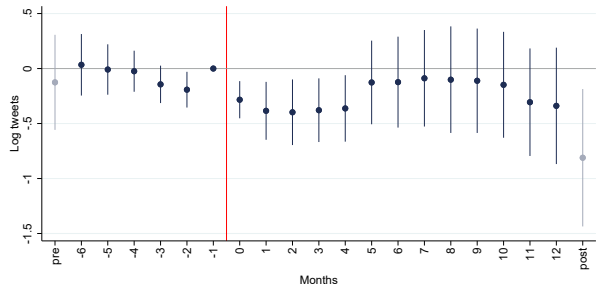
Figure A8: Direct effects of an attack on volume of coverage (alternative specifications)



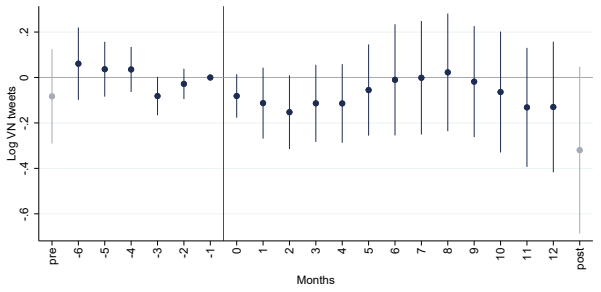
(a) All tweets, month FE



(b) Violence tweets, month FE



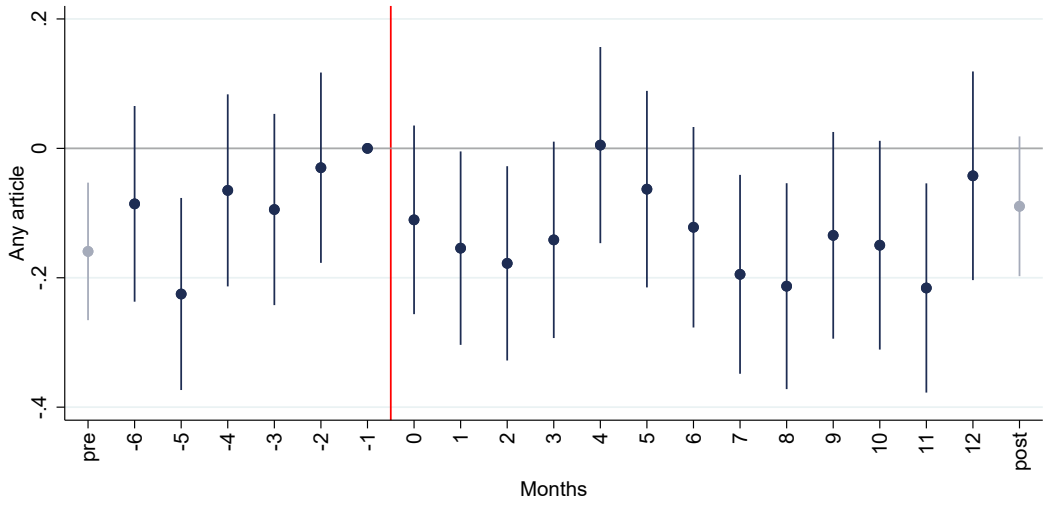
(c) All tweets, month × municipality FE



(d) Violence tweets, month × municipality FE

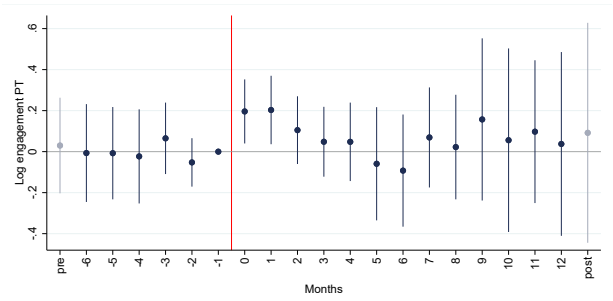
Note: all regressions include outlet fixed effects. Panel *a* and *b* control for calendar month fixed effects. Panel *c* and *d* control for location municipality × calendar month fixed effects. Robust standard errors clustered by outlet.

Figure A9: Direct effects on coverage (Google CSE)

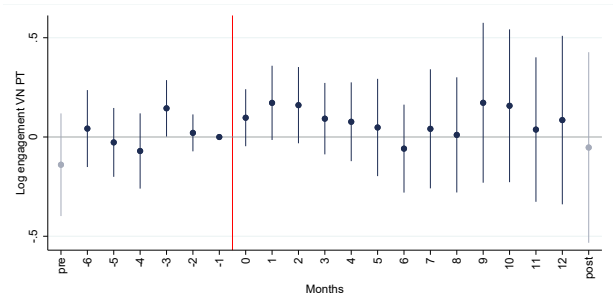


Note: This figure presents event study estimates where the dependent variable is an indicator for *any* articles. Outlet and 30-day period fixed effects. Robust standard errors clustered by outlet.

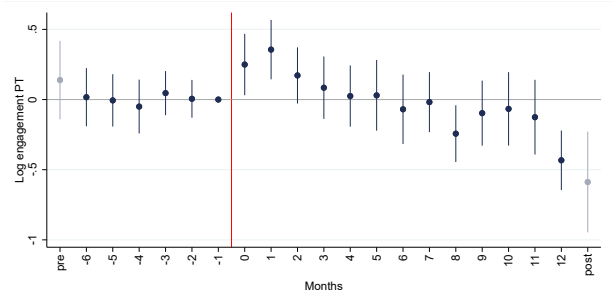
Figure A10: Direct effects of an attack on Twitter engagement for victimized outlets (alternative specifications)



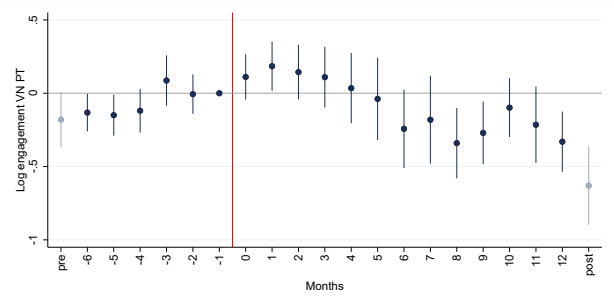
(a) All tweets, month FE



(b) Violence tweets, month FE



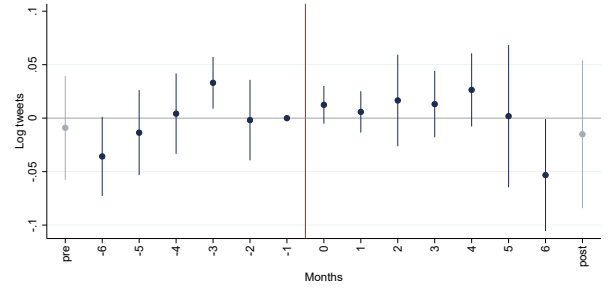
(c) All tweets, month × municipality FE



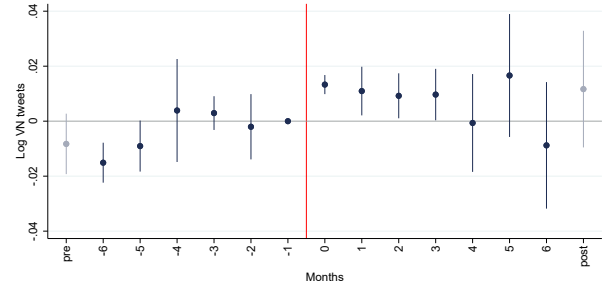
(d) Violence tweets, month × municipality FE

Note: engagement is defined as likes and re-tweets normalized by the number of tweets for a given outlet-day combination. All regressions include outlet fixed effects. Panel *a* and *b* control for calendar month fixed effects. Panel *b* and *c* control for location municipality × calendar month fixed effects. Robust standard errors clustered by outlet.

Figure A11: Indirect effects of an attack on volume, top journalists



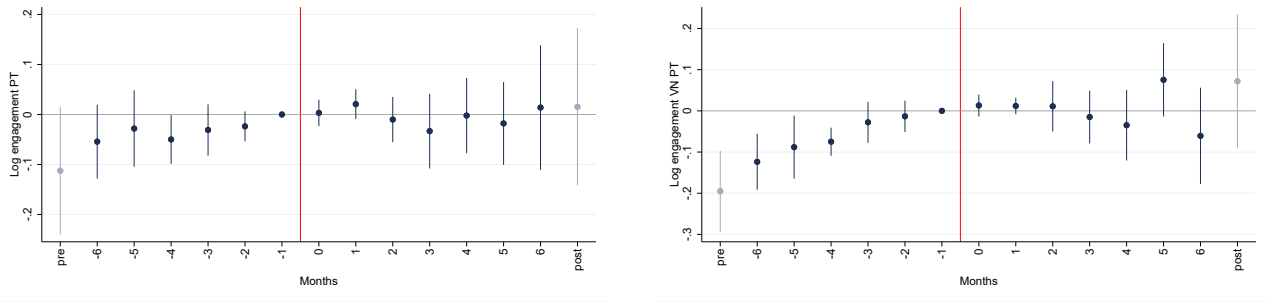
(a) All tweets



(b) Violence tweets

Note: all regressions include journalists × event fixed effects, as well as calendar month fixed effects. Journalists are assigned a state based on volume of keywords. Events are defined by the murder of a journalist at the state level. Robust standard errors clustered by journalist.

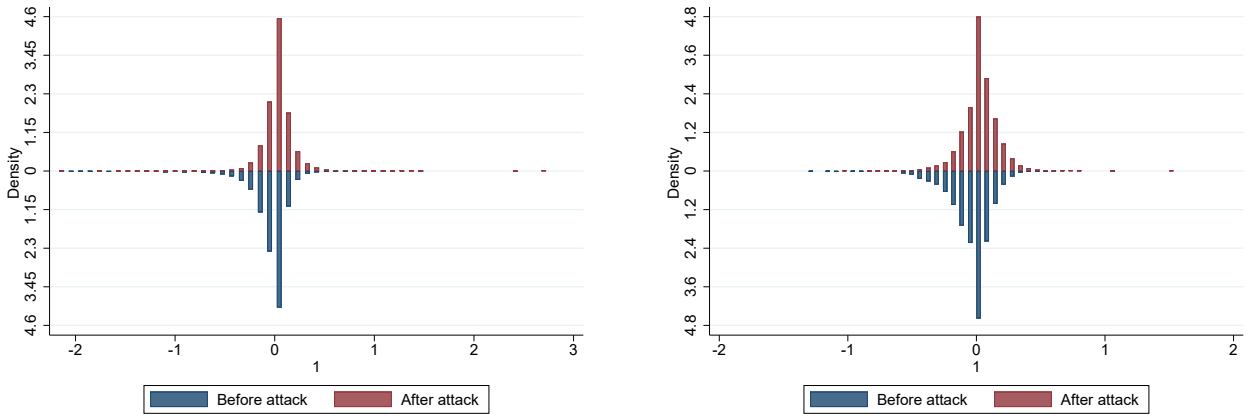
Figure A12: Indirect effects of an attack on engagement, top journalists



(a) All tweets (b) Violence tweets

Note: engagement is defined as likes and re-tweets normalized by the total number of tweets per journalist-day. All regressions include journalists $\times$ event fixed effects, as well as calendar month fixed effects. Journalists are assigned a state based on volume of keywords. Events are defined by the murder of a journalist at the state level. Robust standard errors clustered by journalist. Robust standard errors clustered by journalist.

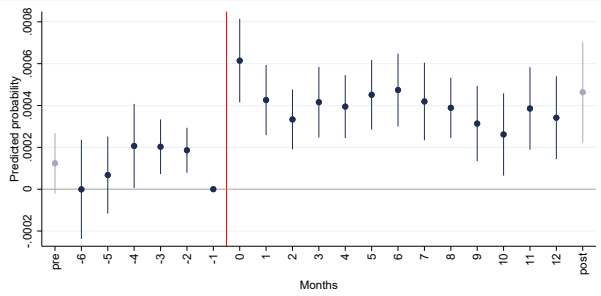
Figure A13: Distribution of Sufficient Reduction by timing of attack



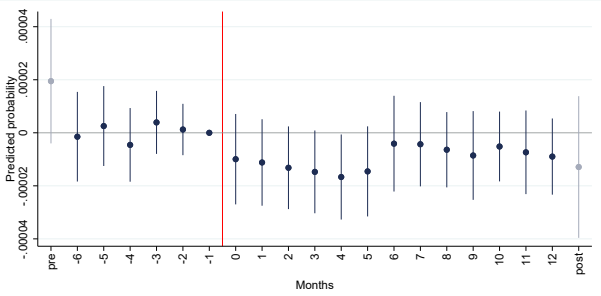
(a) All tweets (b) Violence tweets

Note: depicted is the distribution for the *Sufficient Reduction* statistic (Z) that is constructed through an inverse projection from the MNIR model. The MNIR model is trained to distinguish the timing of a tweet (pre or post-attack) based on text content from tweets 180 days before and after the homicide of a journalist.

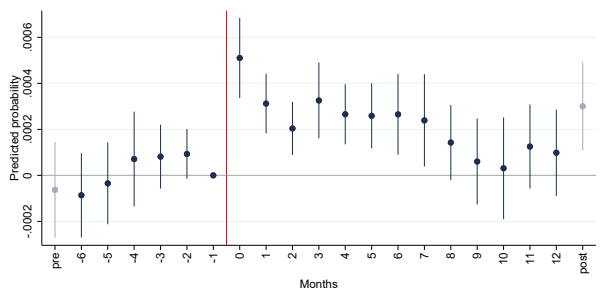
Figure A14: Tone of coverage among targeted outlets (30-day window)



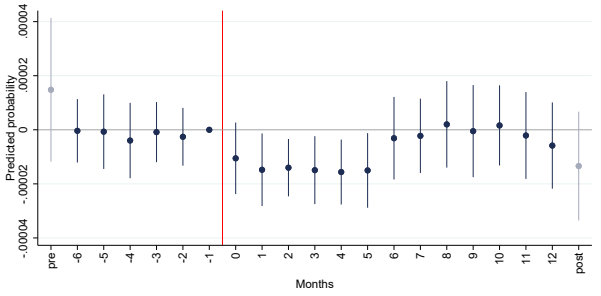
(a) All tweets, calendar month FE



(b) Violence tweets, calendar month FE



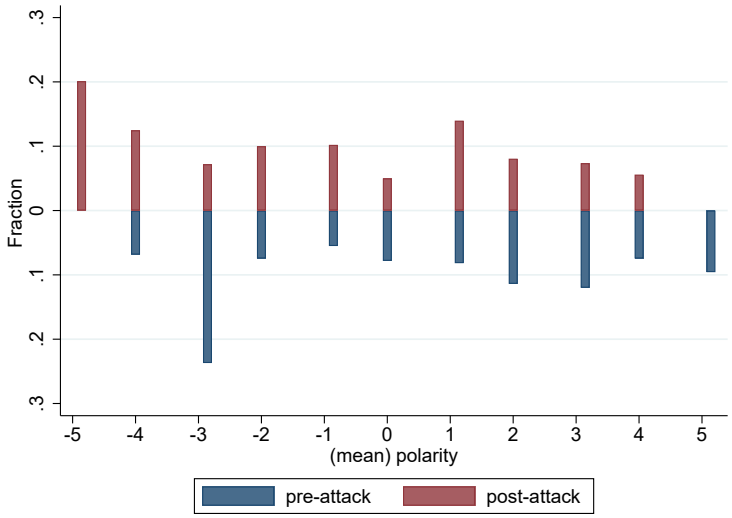
(c) All tweets, calendar month $\times$ state FE



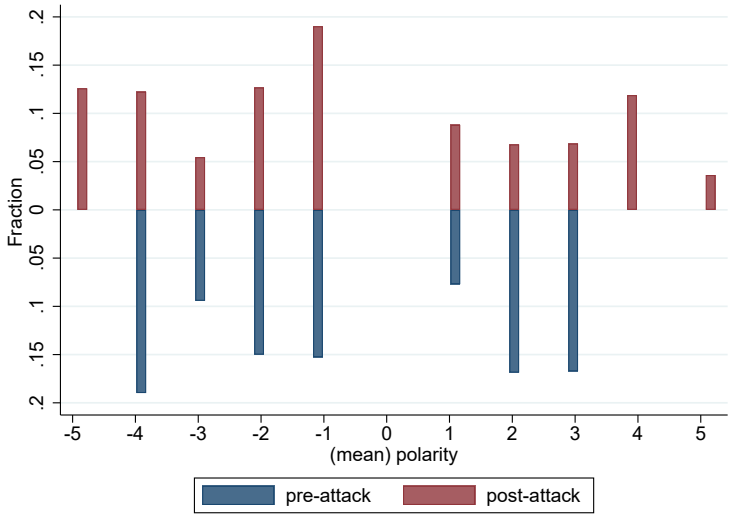
(d) Violence tweets, calendar month $\times$ state FE

Note: we estimate a logit model where we regress the post-attack indicator variable y_{oi} on Z_{oi}, the *Sufficient Reduction* (SR) of the count space of words. This process is referred to as *forward regression* by Taddy (2013). These figures depict event study estimates of the average monthly predicted probabilities of post-attack. Lower values indicate that text content was more similar to the pre-attack content, while higher values suggest that content was more similar to the post-attack content. Both the MNIR model that generates the SR and the logit model are estimated using 30 days worth of tweets before and after a homicide, thus coefficient estimates for $-6, \cdots, -2, 2, \cdots, 12$, as well as *pre*, *post* are out-of-sample. Robust standard errors clustered by outlet.

Figure A15: Distribution of terms by polarity



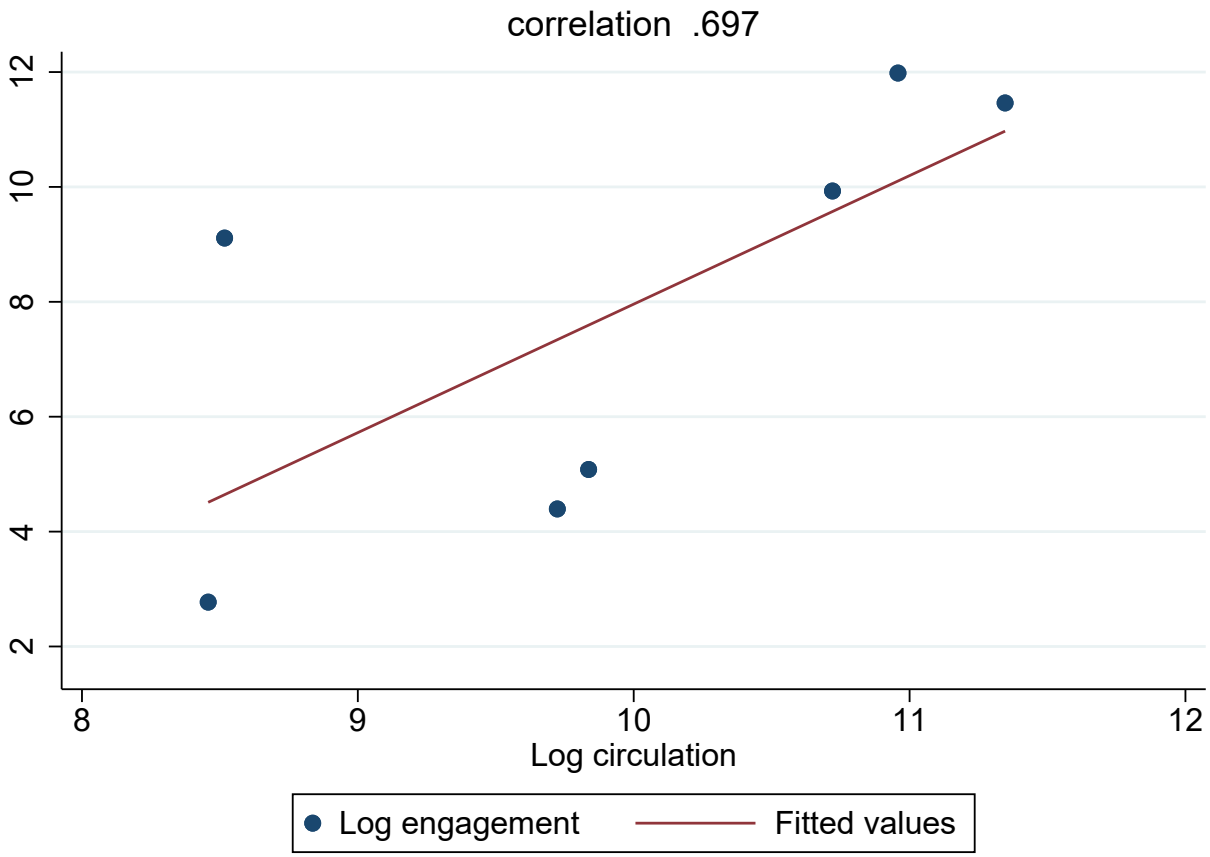
(a) All tweets



(b) Violence tweets

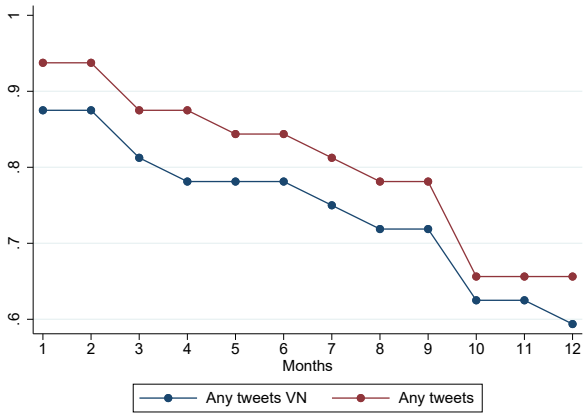
Note: These figures show the distribution of polarity in tweets from targeted outlets. Considered are terms with non-zero coefficient in the MNIR regression 180 days before and after an attack. Polarity is computed from the Brooke et al. (2009) dictionary.

Figure A16: Circulation and Twitter engagement

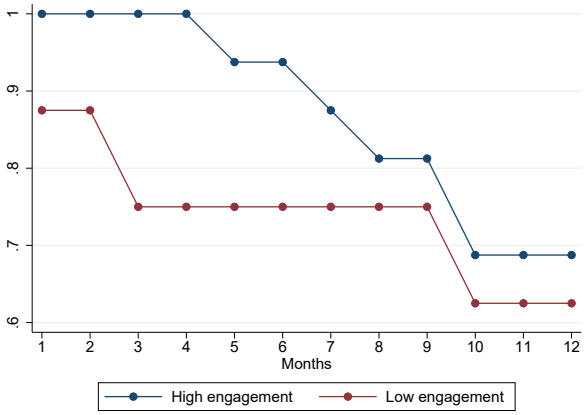


Note: engagement is defined as likes and re-tweets. We calculate engagement from tweets 6 months before an attack on a news outlet, as the event might have an independent effect on engagement. Circulation figures come from the State Secretariat's census on the media.

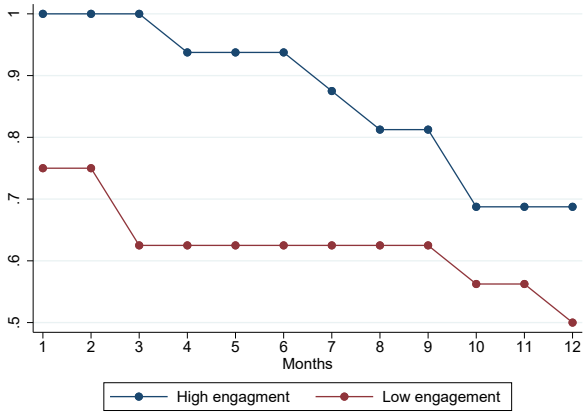
Figure A17: Survival rates of victimized outlets after an attack



(a) All and VN tweets



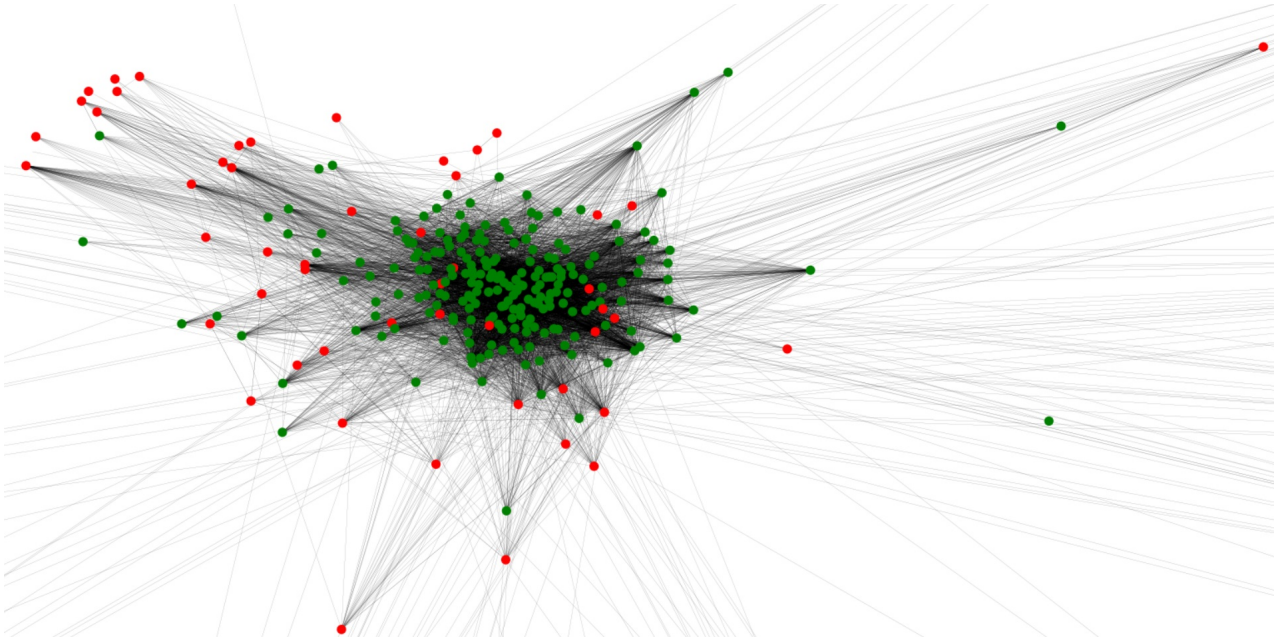
(b) All tweets



(c) VN tweets

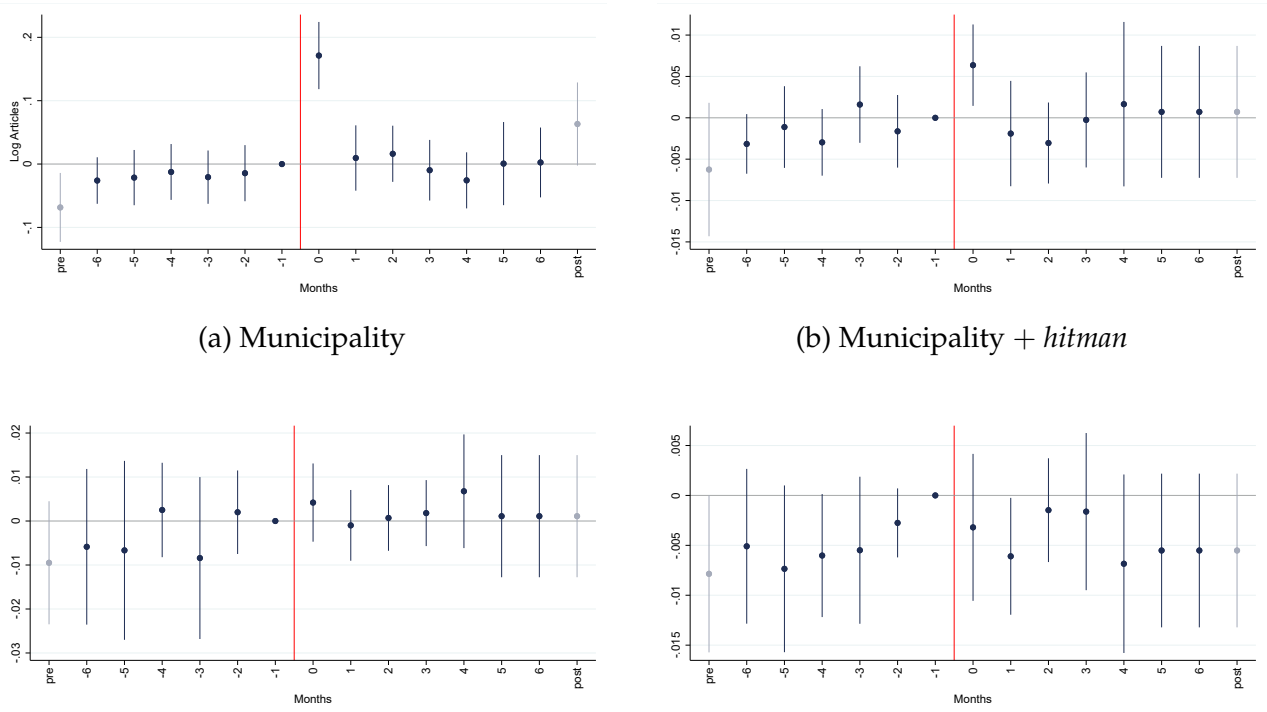
Note: This figure presents the probability that a targeted outlet tweets at least once from month j after an attack through month 12, with $j \in [1, \cdots, 12]$. Only outlets with at least one tweet in the 30 days preceding an attack are considered. Tweets in the day of the attack are ignored, as we want to measure reponses to this event and the time of day when a homicide was initially reported is unknown. *Low engagement* outlets are defined as those with total likes and re-tweets below the median, 6 months before the aggression.

Figure A18: Twitter network of selected accounts



Note: The figure shows a partial Twitter network for the accounts in our dataset. Red nodes represent victimized outlets and green nodes are journalists. An edge is drawn between two nodes if either of the accounts follows the other.

Figure A19: Mentions of municipalities in the national press



(a) Municipality

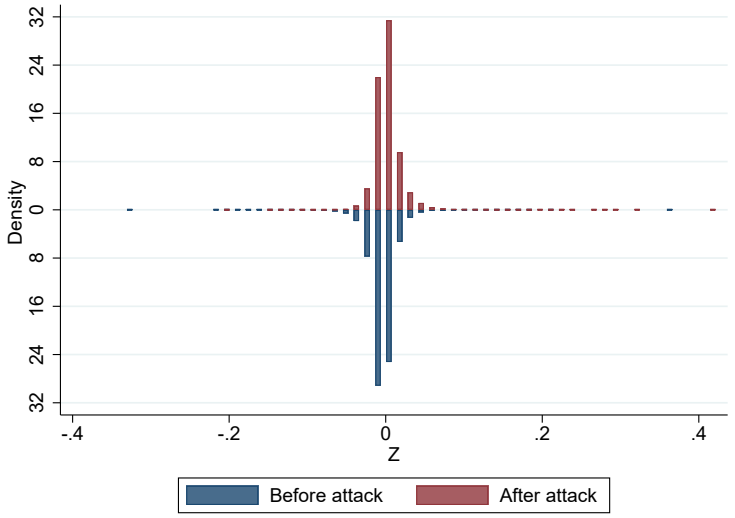
(b) Municipality + *hitman*

(c) Municipality + *Sinaloa Cártel*

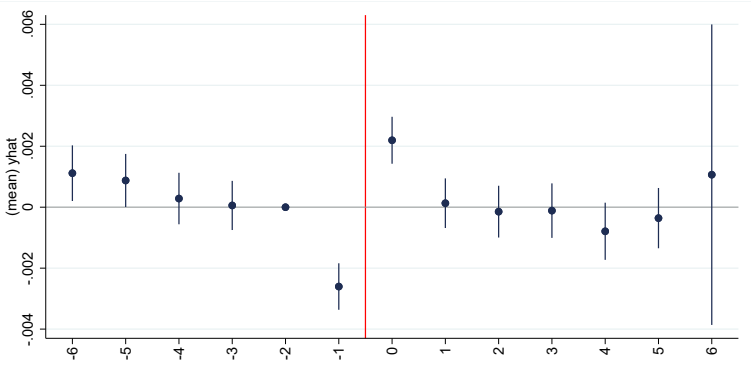
(d) Municipality + *Jalisco Cártel*

Note: event study considers mentions of municipalities in the national press before and after a journalist was killed there. Panel *b* through *d* consider news items that refer to these municipalities, *along with* an additional term. The Sinaloa Cártel and the Jalisco Organization (CJNG) are the largest criminal organizations operating in Mexico and are considered by the US government to be the main criminal threats faced by that country.

Figure A20: Tone of coverage (national press)



(a) Distribution of SR statistic



(b) Probability of post-attack

Note: we train the MNIR model to identify the timing of tweets 30 days before and after an attack. This figures presents the distribution of the Sufficient Reduction statistic (panel *a*). Panel *b* shows the predicted probability of post-attack from a linear probability model where we regress the post-attack status on the SR.