Violence Against Journalists and Freedom of the Press: Evidence from Mexico*

Andres Jurado
Brown University

Juan S. Morales
Collegio Carlo Alberto
University of Turin

October 9, 2020

Abstract

This paper studies how 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 six million tweets published by Mexican journalists and media outlets. We find large reductions in intensity of reporting among victimized outlets following acts of violence. This reduction is consistent with self-censorship, as outlets scale back their publishing while the public, who is interested in these events, engages more with their content following the attack. These murders affect *how* the press reports the news as well: we find subtle but persistent changes in tone of coverage, with outlets that remain active emphasizing the most violent aspects of organized crime after the attack, as opposed to law-enforcement operations. We also document indirect and localized spillover effects on other journalists using a triple-difference design. In particular, after the killings reporting declines among journalists who followed the victimized outlet on Twitter and resided in the same state. Finally, using census data we show that states with more violence against media workers between 2010 and 2015 saw reductions in the number of active journalists and changes in their demographics.

*We thank Brian Knight, Pedro Dal Bó, Jesse Shapiro, and seminar and conference participants at Brown University, the Collegio Carlo Alberto, University of Pittsburgh, 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 democracy. Yet, 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 violence against the media and news reporting in Mexico, a country suffering from a high level of drug-trafficking related violence and characterized by one of the highest murder rates in the world (as well as accounting for 5 percent of all journalists' killings).

Despite the frequency of targeted violence against journalists, systematic studies of its effects on news reporting remain scarce. Qualitative evidence suggests that criminal organizations' harassment of and violence against media workers[2] could have mixed effects on reporting. On the one hand, violence may deter the press from covering certain events. A journalist working along the US border reports:

> "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 acts of violence may cause backlash from journalists:

> "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 do it"
> (Relly and González de Bustamante, 2014, p. 116)

We study the relationship between violence against media workers and news reporting using data on all reported killings of members of the press in Mexico between 2006 and 2017. We combine this with newly compiled data of news articles and press activity from both online media sources, including Twitter and print media, as collected in Eficiencia Informativa, a private database with over thirty-five million news articles from Mexican news outlets. Our Twitter dataset consists of seven million tweets published by more than

[1] Committee to Protect Journalists, 2019 (https://cpj.org/), accessed April 9, 2020.
[2] We use the terms "media worker", "journalist" and "reporter" interchangeably.

three hundred media accounts between 2009 and 2017. These include 226 popular Mexican journalists and seventy-six outlets that lost one of their employees to homicide. In addition, we use data from the Mexican census to study long-term changes in the demographics of the journalist profession.

We exploit the precise timing of journalists' killings in a series of event-study exercises, through which we measure short and medium-run changes in news activity following the violent incidents. We differentiate between the direct effects of being a victim of targeted violence and the indirect (spillover) effects on nonvictimized outlets and journalists.

Our analysis reveals sustained reductions (around 25 percent) in victimized outlets' Twitter activity following the murder of an employee. Smaller outlets are especially sensitive to these attacks, with 20 percent of them exiting the market for crime news permanently. This could be explained by losing a larger proportion of their human capital to an attack, which we refer to as a "mechanical effect". We also document indirect effects that are highly localized and concentrated on journalists who are likely most at risk: those in the same Twitter network, defined by whose accounts they follow, and state as the victim.

We use a modern supervised machine-learning algorithm to quantify changes in tone among victimized outlets following an attack, and we find persistent effects. We show that words that emphasize the violence of organized crime appear more frequently afterwards, whereas prior to the attacks we find mostly words that describe law enforcement operations. We also employ a third-party polarity dictionary and find that language becomes more negative following the homicide. While this evidence is in line with the hypothesis that outlets that remain active exhibit backlash, these effects are also subtle and relatively muted.

We complement our analysis using data from the Mexican census. Since 2010, the census and the Inter-Census Survey have included questions that allow us to identify the number of journalists active in the thirty-two states in the country. We implement a difference-in-differences estimator to measure long-term effects of violence against media workers.

States with more homicides have fewer total active journalists than they otherwise would have. Press demographics also appear to change as a result of this, with journalists earning lower wages and being less likely to be married and have kids than before the rise in violence.

Our work is related to the literature on the political economy of the determinants and effects of media coverage. The importance of the 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) find that exposure to Fox News increases voting for the Republican Party. Nevertheless, other evidence finds limits to the influence of news media: consumers tend to discount information from sources they consider biased (Chiang and Knight, 2011), and behavioral mechanisms can limit the influence of the press (Durante and Knight, 2012; Knight and Tribin, 2019).

Closer to our work, several papers have studied sources of media bias. di Tella and Franceschelli (2011) show that public funds spent on advertising affect how the press covers negative events involving the government, and Beattie, Durante and Knight (2017) and Reuter and Zitzewitz (2006) show that private financial incentives can change how the press covers the news. Gentzkow and Shapiro (2010) document that newspapers in the United States respond to a public preference for news that reflects their opinions by choosing to slant in a profit-maximizing manner. In the context of Mexico, Ramírez-Álvarez (2020) reveals how editorial standards and informal agreements between publishing houses and the government influenced how the press covered the War on Drugs.

Previous studies have mostly focused on state-sanctioned censorship in authoritarian environments. King, Pan and Roberts (2013), Qin, Strömberg and Wu (2017), and Roberts (2018) have documented how the Chinese government faces a trade-off between censorship and the ability to monitor collective action in online media, while Qin, Strömberg and

Wu (2018) explore the trade-off between political and economic interests in print media. Imprisonment of dissident leaders is another common tool employed by autocratic states, yet its effectiveness in Saudi Arabia appears limited: the government is able to silence dissidents, but spillover effects to followers are meager (Pan and Siegel, 2019). Our work instead considers non-state sanctioned censorship in an electoral democracy.

A number of papers, most notably in the fields of political science and journalism, have studied repression of journalists in Mexico. 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 predict targeted media violence, and Hughes and Márquez-Ramírez (2018) document that being in an environment with higher common 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 critical of the government, the presence of newspaper networks and NGOs can mitigate that effect.

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 killings, 11 journalists were kidnapped before being murdered and 9 were tortured in addition. 84 percent of homicides were classified by CPJ as "murders", while the rest were attributed to "dangerous assignments." In 79 percent of cases CPJ attributed the attack to criminal groups, followed by government and the military (7.9 percent each) and local residents (5.3 percent). At least 22 journalists received threats (*ibid*).

Criminal organizations often try to influence news reporting, and suggest that these murders were punishment for 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."[3]
>
> (Valdez, 2016, p. 42)

Figure 1 presents the proportion of victims that reported on various popular subjects, based detailed accounts by CPJ: 78 percent covered crime or the police beat, 34 percent politics, and 20 percent corruption (some journalists cover multiple subjects, so the figures do not add up to 100 percent).

The locations of these murders indicate connections to organized crime: states that are important points of entry for drugs tend to experience more murders of media workers. Figure A1 depicts murders of journalists and homicides among the general population for the thirty-two states in the country. Journalists are especially at risk in the states of Guerrero, Oaxaca, and Veracruz (dark blue), which together account for 40 percent of all murders, but only for 12.8 percent of the country's population. Guerrero and Veracruz are notorious as ports of entry for cocaine imported from Colombia, and for producing other illegal narcotics, while Oaxaca suffers occasional outbreaks of ethnic conflict. The eastern state of Veracruz is by far the most hostile environment for the press, with 18 percent of all media workers killed. Attacks on members of the press are not a simple function of overall violence, as can be seen by the weak spatial correlation between total number of homicides and murders of members of the press. Northern states, especially those at the border with the United States, experience high homicide rates but report comparatively few killings of journalists.

These high level of violence is a recent phenomenon and is linked to the start of the Mexican War on Drugs. Figure A2 shows how the targeting of media workers increased

[3] Own translation.

steeply in 2006, followed two years later by an increase in homicides among the general population. The year 2006 marks the last year of the war between the criminal syndicate called the Federation and some of the Gulf of Mexico-based cartels. With a cease-fire in place in 2007, the number of media workers murdered fell, as did total homicides in the country. By 2008, President Calderón's militarized strategy against drug-trafficking organizations was underway; the Federation splintered into two bands fighting for control of profitable drug-trafficking routes. This coincides with the increase in killings of media workers and members of the public that persists to this day.

Perhaps one reason behind this high level of victimization is the low number of cases that are solved by the police. CPJ reports that in 86.8 percent of cases there was complete impunity, and in 10.5 percent partial impunity. Justice was served only in 2.7 percent of cases. In a 2017 report, it concludes:[4]

> 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 CPJ are two leading NGOs that advocate for journalists. Both keep track of murders of media workers in Mexico, along with the workers' affiliations. Out of 104 victims, 17 were classified as free-lancers and hence are not matched to any outlet. The rest are assigned to 103 outlets (some journalists had more than one affiliation).

Both of these organizations 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 body was later found. In those cases, we follow the convention of using the earliest plausible date.

[4]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.

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. Its archive contains over thirty-five million full news articles, tweets, rand adio and TV transcripts. The company is certified ISO 9000 for quality management.

We collected a sample of 2.2 million news items by including items that match any of the following keywords: "narc*", "sicari*" (hitman), "difund*" (spread, as in "spread information"), "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 in Ramírez Alvarez (2017); the regression predicts whether an article is related to drug-trafficking. We included in our list additional terms that we believe also describe crime-related events. These articles range from January 2006 to October 15, 2017.

Of the items in the dataset, 22.5 percent are transcripts of radio broadcasts, 16.5 percent are transcripts of TV broadcasts, 0.2 percent correspond to the *Official Gazette of the Federation* (*Diario Oficial de la Federación*), and the rest, 60.8 percent, are from print media. EFIC tends to track outlets with high circulation, as these are more likely to be of interest to its target audience: the government, private individuals and corporations. While our database contains 169 outlets, just 2 leading newspapers, *La Jornada* and *Reforma*, account for 248,000 articles, or 11.2 percent of the total.

Our sample from EFIC contains many terms related to drug violence, such as "security", "federal_attorney", "investigate", "police", "organization", "seize", "dead", "army", "crime", "law", as can be seen in figure A3. Figure A4 panel a plots the monthly number of articles in EFIC against total homicides in the country. These series exhibit a 0.82 correlation coefficient.[5] The EFIC database thus seems to deliver high-quality data with which

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

to investigate our questions of interest.

3.3 Twitter

We built two distinct datasets of tweets using a selected set of usernames. The first dataset includes tweets from 224 journalists whose usernames were collected from Twitter-Mexico.com, a website that archived and documented popular Twitter users in Mexico.[6] The second set of usernames corresponds to seventy-six victimized outlets (as documented by the CPJ and Article 19), which we were able to manually match to corresponding Twitter accounts.

We then used Twitter's Advanced Search tool 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 users often do not tweet more than twenty times in one day, this limit is not always binding.[7] 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 is binding for 17 percent of observations in the journalists dataset and for 42 percent of observations in 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 outlets dataset.[8] Figure A4 panel b shows the correlation coefficient between number of tweets with violent content by the top journalists and number of homicides in the country is 0.43. This correlation is likely underestimated because of the 20-tweet limit, and an al-

[6] Though the site is no longer online, the web page we used can be accessed through the Wayback Machine.

[7] 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.

[8] The set of words is the following: *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.

most five-fold increase in tweets in December 2014, which was likely caused by coverage of the massacre of Ayotzinapa (see footnote 5).

3.4 Other data

We gathered data on daily-circulation of thirty-five victimized outlets from the State Secretariat's National Census of Printed Media[9] (Padrón Nacional de Medios Impresos). The Census of Printed Media also records the municipalities where outlets are distributed. These figures are collected by third-party auditors paid by the outlets, which might prove too onerous for smaller firms. Hence, larger outlets tend to be over-represented.

General-population homicides are reported by the National Statistics Institute. We use daily counts of homicides at the municipality and state levels. The Centro de Investigación y Docencia Económicas (CIDE) Drug Policy Program (PPD) maintains a database with homicides attributed to drug-trafficking organization compiled by a government panel. We decided against using it, as the Mexican government ceased to tracking murders of this kind in 2011.

Finally, we access data on the 2010 and 2015 Mexican censuses through IPUMS International.[10] The analyzed sample is restricted to journalists and occupational categories in the census that are near to journalists; the categories include accountants, researchers, psychologists, and artists and performers. Summary statistics for the journalists in our sample are in table A1, separately for states where at least one journalist was murdered and those where none was murdered.

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

[10] We do not use previous censuses, as we can not identify journalists as an occupational category before 2010.

4 Empirical Analysis

Different mechanisms, acting in opposite directions, might shape the media's response to an attack. In characterizing these mechanics, we adhere to previous work when appropriate (Pan and Siegel, 2019). Volume of coverage might decrease simply because fewer journalists are reporting for a given outlet, which we refer to as a *mechanical* effect. The pool of news to report now includes a high-profile murder (*content* effect). Outlets may change their publishing behavior (*behavior effect*) in response to the attack. More specifically, the objective of the criminals is likely to deter future publishing of certain events through fear. Outlets might 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, as journalists might step up their publishing to signal that they will not be intimidated or to protest the murder of a colleague (*backlash* effect). The attack itself might increase the public's interest in the content from targeted outlets, which could lead to more intense publishing activity (*demand* effect). Lastly, these mechanisms may work through the targeted outlet (*direct* effects), or by other journalists (*indirect* effects).

In section 4.2 we show that attacks on journalists reduced the volume of tweets by targeted outlets, while in section 4.3 we show that engagement (likes and re-tweets) and Google searches for homicides of journalists increased substantially. Demand forces and content effects, it appears, are of secondary importance, and the combination of mechanical and deterrence effects 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 are concentrated among victimized outlets; spillover effects within the same state as victimized outlets 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 more attacks 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. The evidence is consistent with mechanical effects being the main drivers of censorship; but it is also consistent with behavioral effects if media workers in outlets that were not victimized, even after accounting for new information, believed that their coverage was unlikely to place them in danger. We find suggestive evidence for both effects. In section 4.2 we show that smaller outlets exited the market for news at greater rates than larger outlets following an attack, which is consistent with mechanical effects. In section 4.5 we show that journalists most at risk reduced their Twitter activity compared to those that we estimate were less at risk, which is evidence of indirect deterrence. Finally, in section 4.4 we discuss how text content became more negative after acts of aggression, with words describing the most violent aspects of organized crime appearing more often, which is consistent with a backlash effect.

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, first, the existence of other newsworthy events that may affect coverage directly by changing the pool of news to report and, second, the expectation of an attack, which 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 was an unusually high number of homicides among the general population[11] leading to the act of aggression. Murders are newsworthy events and constitute for many outlets one of the sections that most drive sales. An increasing number of murders may also indicate increased competition among criminal groups which could

[11] We consider murders *net* of homicides of journalists. Of the approximately thirty-three thousand annual homicides per year in the country only a dozen correspond to journalists, on average.

make the job of the press riskier, as these groups 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 attack. The model we estimate is as follows:

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

$homicides_{set}$ is the log of homicides (net of press homicides) 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 thirty-day with respect to the attack, and Pre_{set} and $Post_{set}$ are binary variables equal to 1 for $t < 6$ and $t > 6$, respectively.[12] Many states experienced more than one homicide of a media worker. In those cases, we pair each monthly window to the closest event and define time windows relative to the event. 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 media worker homicide *occurred* and the municipality where targeted outlets are *located*. States with a media worker killing reported an increase of 2.5 percent in homicides compared with the preceding month, but this difference is not statistically significant. The municipalities where victimized outlets are located reported virtually the same number of homicides from one month to the next. 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 the hypothesis that journalists affiliated with victimized outlets did not anticipate the attacks.

[12]Time windows are defined based on the day of the attack, such that attacks are coded as taking place during the first day of time window 0. We chose twelve-month windows because homicides are clustered in time in states where the media is victimized, which limits our ability to identify coefficients far from the event date. A stark example is the state of Veracruz, where we observe three consecutive months with an act of aggression.

CPJ, for instance, reports known threats to only 16 victims out of 104, and some victims appear to have disregarded the threat.[13]

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, at time t. We include as a control the log of homicides in past thirty days in the municipality where the outlet is located, 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 fixed effects for state $\times$ month (γ_{st}) and outlet fixed effects (γ_o). The coefficients of interest, β_k, capture the change in Twitter activity of victimized outlets relative to the activity of other outlets in the same state that 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, although in the short run only the coefficients for the full sample are statistically significant. Publishing falls to its lowest point around three to four months after the act of aggression, which might be explained by initial reporting about the attack partially offsetting a later reduction in activity.

[13] "[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.)

One concern with our main estimate is that another event that affects volume of coverage might 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. These controls account for all local events at the municipality level, such as local elections, that might affect our results. Reassuringly, coefficient estimates are little affected. These finer fixed effects control for local unobserved factors but reduce our sample size to only states and municipalities with more than one victimized outlet.[14] Estimates including calendar-month FE (as opposed to fixed effects for state $\times$ calendar-month) are also in line with the baseline results, which suggests that spillover effects on nonvictimized outlets within the same state may be limited.

We have documented how violence against media workers 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 using Google Custom Search Engine (CSE) API. Therefore, results using Twitter are likely to be a good proxy for the effect on newspaper articles.

We investigate heterogeneity in the effects by outlet size. If smaller outlets are disproportionately affected by the attacks, this might indicate the presence of mechanical effects: losing a reporter might hinder the reporting ability of smaller outlets more than it does for larger outlets. Unfortunately, we do not have access to the number of employees of news organizations. While circulation figures might be a good measure of size, we have these figures (from the State Secretariat's census on the media) for only a handful of victimized outlets. Instead we consider Twitter engagement data, readily available, as a proxy for circulation. We find that engagement is indeed a good proxy of circulation for the subsample of outlets for which we have these data (figure A15).

A natural way to frame the question of heterogeneity is in terms of survival rates for

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

outlets following an attack. We consider outlets with at least one violence tweet in the thirty days preceding an act of aggression and compute the probability that outlet o tweeted at least once between month j and month 12 following an attack, $Pr(\sum_{k=j}^{12} tweets_{ok} > 0), j \in [1, \cdots, 12]$.

Panel a of figure A16 shows that within a month after an attack 5 percent of outlets ceased to be active on Twitter and 12 percent no longer tweeted news about violence. The number of active outlets drops by another 7 percent between months 2 and 3. Twelve months after an attack, 35 percent of outlets were no longer active, and 40 percent no longer tweeted news about violence. Panel b further breaks down survival rates by whether the outlet has low or engagement, which we define based on the median likes and re-tweets six months before an act of aggression. While initially the same percentage of low- and high-engagement outlets survived, 15 percent of low-engagement outlets exited the market between months 2 and 3. Panel c shows that while 5 percent of high-engagement outlets ceased to tweet news about violence one month after an attack, the equivalent figure for low-engagement outlets is closer to 20 percent. Another 10 percent of low-engagement outlets ceased 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 Public interest and attacks against the press

This section studies the reaction of the public to an act of aggression against the press, as this may have an independent effect on publishing. Changes in demand for the content of victimized outlets following an attack constitute an important mechanism through which outlet publishing may change. For instance, increased demand for the outlet's content (which can result form increased notoriety) might incentivize the press to increase their reporting activity. On the other hand, if the victim of homicide was producing content that resulted in sales for the newspaper, his or her demise might decrease public demand for content.

We test whether the public's interest in victimized outlets' content changed by regressing engagement per tweet on the timing of an attack, in a similar specification to that in model 2. Engagement is defined as the log of likes and retweets normalized by the number of 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 a slightly smaller increase for tweets about violence). Results are similar when considering alternative 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 change 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 than before. Nevertheless, these changes in tone are relatively muted, to the point that 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 after a 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 ten thousand newspapers.

Lastly, we do not find any evidence of changes in engagement with journalists located in states where the attacks occurred (Figure A12), even though there are 2 percent 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 the homicides, as opposed to changes in reporting.

4.4 Direct effects of violence on tone of coverage

An attack on the press might affect not just the volume of news, but also how events are covered. These changes might include both the types of news and the language used. We loosely refer to both features as "tone," following the 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 act of 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 for tweets with violent content. Because the analysis relies on a medium that follows language conventions loosely, we aggressively clean up the text. First, we process terms through a so-called text pipeline in which we drop common prepositions and articles, as these convey minimal useful information. We also lowercase words and eliminate symbols, numbers, and hashtags to reduce the feature set to consider. We 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 versions of words are conceptually similar and to limit the impact of certain types of orthographic errors. We further avoid including names and words with orthographic errors by only considering stemmed terms from the Spanish version of the dictionary published by the Royal Academy for the Spanish Language (Real Academia de la Lengua Española).[15]

[15]To the best of our knowledge there is no machine-readable dictionary for the Mexican version of the

Lastly, we consider only terms that appear in at least thirty tweets to limit overfitting and prevent some orthographic errors.[16] 2,933 terms meet this criteria for the entire sample of tweets (and hence are considered in the analysis), as well as 729 terms for sample of violence tweets.

We train the MNIR model on this set of words and consider the set of terms with nonzero coefficients, which is approximately one-third of all terms. Importantly, we ignore tweets about murdered media workers (approximately 1 percent of all tweets), as the press covered these events extensively and hence the resulting set of terms picked by the MNIR model is not informative. Specifically, we filter out tweets that either 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 timing, but low values are less indicative of actual timing. This may indicate that new "themes" appeared after the attacks. For violence tweets (b) the two distributions are more similar but with the post-attack distribution slightly shifted to the right. Here, high values of Z are indicative of the post-attack period and low values indicative of the pre-attack period. Nevertheless, we see a fair amount of overlap in the distributions. Thus, any changes in tone are likely subtle, which could be because the sample is more homogeneous.

We formally test for changes in tone by predicting the probability that a tweet was

Royal Academy for the Spanish Language.

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

published after an 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 MNIR model assigns a 0.05 percent higher probability of post-attack timing to tweets that were indeed published after an attack, relative to those published before. 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 fixed effects for calendar month $\times$ state.[17] Importantly, the outlet fixed effects ensure that we capture *within*-outlet changes in language, as opposed to any compositional changes that might result from certain outlets reducing their coverage more than others. Lastly, while overfitting 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.

To explore how content itself changed, we consider the set of words with nonzero 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 by regressing our indicator variable of post-attack timing, y_{oi}, on the within-tweet fraction of terms with nonzero loading, $\hat{f_{oi}}$. Figure 6 presents the distribution of loadings and frequencies for the full sample of tweets and the subsample of violence tweets. The top thirty terms that most contribute to identifying the timing of tweet are highlighted, where contribution is defined as the product of loading and frequency.

Panel a shows that words with large contributions to predicting pre-attack timing among the full sample are mostly unrelated to violent events. Such words include "news," "state," and "north" (probably referencing popular northern music). Terms that likely ref-

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

erence violent events include "vehicle," which is often mentioned in the context of seizures or arrests carried by law enforcement. There are more such terms after an attack: "execution_or_perform," "dead," "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 connotations as predictors of timing.

Panel b limits the impact of an increasing share of violence tweets post-attack by presenting results for the sample of tweets with violent content. Terms that predict pre-attack timing are closely related to crime with surviving features such as "alleged," "crime," "war," "army," and "agency." Thus, it appears that outlets report more about crime and about 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."

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 effects on tone of coverage by studying changes in polarity as a result of an act of 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 crowdsourced classification via Mechanical Turk. Other methods tested, by the authors such as vector machine learning and automatic translation of an existing English dictionary, performed worse than this baseline dictionary, which is 74.50 percent

accurate. Polarity is coded 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 -0.05 to -0.1 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. These patterns are starker when controlling for fixed effects for calendar month $\times$ state, which implies that targeted outlets experienced a greater decrease in polarity compared to outlets in the same state. Figure A14 shows the distribution of terms by polarity before and after an act of aggression. In panel a, 10 percent of all classified terms in the full sample of tweets had a polarity of 5 (the maximum possible) before the event and a similar number with polarities 3 and 4. Approximately 8 percent of terms had a polarity of -4, the lowest observed. Following an attack, 20 percent 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. Afterward, 12 percent of terms exhibited very negative polarity (-5) but 4 percent had very positive polarity (5). This likely explains the initial muted change among violence tweets.

4.5 Indirect effects

In this section we examine attacks' indirect effects on nonvictimized journalists. In particular, we examine whether there were localized spillover effects by examining a narrow set in terms of both physical and social distance from the killing: journalists in the same state as the victimized journalist and in the latter's Twitter network.[18] If there were indirect behavioral responses, we might observe either an increase in tweets denouncing the act of 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 (Halber-

[18] Figure A17 shows the Twitter network for the accounts in our dataset.

stam and Knight, 2016) and to cover similar content. Accordingly, journalists who followed the victimized outlet on Twitter may have responded to the killing differently from other journalists. In figure 4 we compare the behavioral response of journalists in the victim's network *and* state to the response of other journalists. The model we estimate is as follows:

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

y_{jsft} is the log of 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 fixed effects for state $\times$ month (λ_{sm}) and fixed effects for victimized-outlet follower $\times$ month (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 relative to followers of the victimized outlet who are located in different states. We observe a sustained reduction of around 10 percent in Twitter activity starting in the second month after the event (panel a) in the short to medium run and a larger reduction in the long run. Though the estimates are noisier, there appears to also be a reduction of around 5 percent in coverage of violence (panel b). The results suggest that the killings are effective in reducing coverage not only by the victimized outlets but also by their nearby peers.

4.6 Long-run effects

The previous sections showed that targeted outlets reduced coverage in the short to medium run, while spillover effects were modest. Attacks against the press might, 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 on the Mexican census, which started recording data on individuals working as journalists in 2010. We show that more-dangerous states experienced greater 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, and 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 more violence against the press is associated with changes in the share of journalists. In particular, we rely on regressions of the following form:

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

The outcome of interest, y_{ist}, is an indicator equal to 1 if individual i in state s in 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 (alternatively, an indicator equal to 1 if at least one journalist was killed in the state), and the $post_t$ indicator equals 1 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 more between 2010 and 2015 in more violent states. Because about 3.5 percent of workers in the 2010 sample worked for the press, this implies that in states with one or more killings, the share of journalists decreased by 25 percent ($\beta = -0.0087$, column 4) relative to states where no killings took place (and relative 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 changed as well. We rely for this on the following triple-differences regression model:

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

The outcomes of interest, y_{isto}, include demographic and labor-market characteristics, such as number of children, marriage status, years of education, age, and income, among others, for individual i in state s in 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 more media workers killed, journalists in 2015 are less likely to be married, are likely to have fewer children, are less likely to 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 media workers. We document that these events received plenty of attention and that the press generally did not attribute these attacks to any criminal organization, either out of lack of knowledge

or for fear of reprisal. We also show that the acts of aggression did not lead to permanent changes in reporting, in line with what we might expect because large national outlets are not generally targeted by criminal organizations.

Figure A18 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[19], which is approximately 90 percent of them (municipalities are often named in Náhuatl and other indigenous languages, and thus we are unlikely to misclassify 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 was a 0.17 log-point increase in mentions of the municipality. The effect subsides 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."[20] 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 (panels c and d).

Figure A19 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 within thirty days of an attack in a given state. Panel a shows that some post-attack news items have high values of SR but some have very low values. It also shows a high degree of overlap in the two distributions around zero, which means there is 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 timing based on the SR. There is an approximately 0.4 percent jump in probability after an attack, but

[19]Some municipalities share names with one or more other municipalities or with commonly used nouns.

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

this effect quickly subsides (for ease of interpretation, we normalized the coefficients with respect to $t = -2$). The fact that coefficients from 2 through 6 are all close to 0 suggests either that there are no persistent effects in tone or (more likely) that the jump in probability that we measured was driven by overfitting in the MNIR model and thus that tone may have not changed 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 threefold. This paper studies how the news media in Mexico changed its coverage of violence in response to the killings of journalists and other media workers.

We document that a large majority of victims had covered crime before they were attacked. While 97 percent of these murders remain unsolved, press reports suggest that drug-trafficking organizations planned and carried out these homicides. Victims were affiliated with 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 causing censorship. Smaller outlets, which lost a higher share of their employees to attacks, were especially at risk, with fully 20 percent 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, or live in urban areas and earned lower income, which could indicate lower risk aversion.

Our measured effects are the result of not only mechanical effects but also two 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 violent aspects of organized crime, such as extrajudicial executions and confrontations. In doing so, polarity became more negative. Second, journalists in the same network and state as the victim (thus likely more at risk than other journalists) reduced their Twitter activity comparatively.

This paper contributes to the literature by documenting how effective violence is at censoring the press in the context of Mexico and, in doing so, finds yet another externality of drug trafficking. Limiting the flow of information, particularly local information that might not be reported elsewhere, hurts 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.

King, Gary, Jennifer Pan, and Margaret E Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American Political Science Review*, 326–343.

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.

Qin, Bei, David Strömberg, and Yanhui Wu. 2017. "Why does China allow freer social media? Protests versus surveillance and propaganda." *Journal of Economic Perspectives*, 31(1): 117–40.

Qin, Bei, David Strömberg, and Yanhui Wu. 2018. "Media bias in China." *American Economic Review*, 108(9): 2442–76.

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

Ramírez-Álvarez, Aurora Alejandra. 2020. "Media and Crime Perceptions: Evidence from Mexico." *The Journal of Law, Economics, and Organization*. ewaa010.

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.

Roberts, Margaret E. 2018. *Censored: distraction and diversion inside China's Great Firewall.* Princeton University Press.

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.

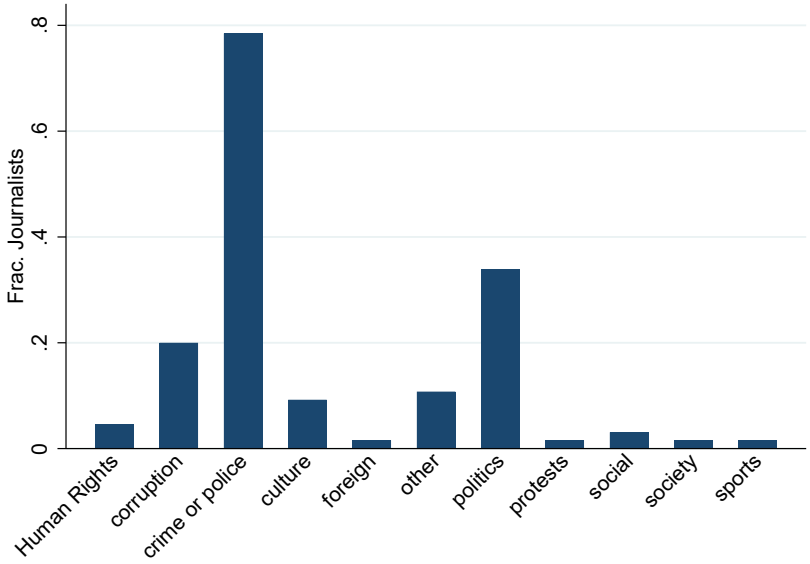
Stanig, Piero. 2015. "Regulation of speech and media coverage of corruption: An empirical analysis of the Mexican Press." *American Journal of Political Science*, 59(1): 175–193.

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.

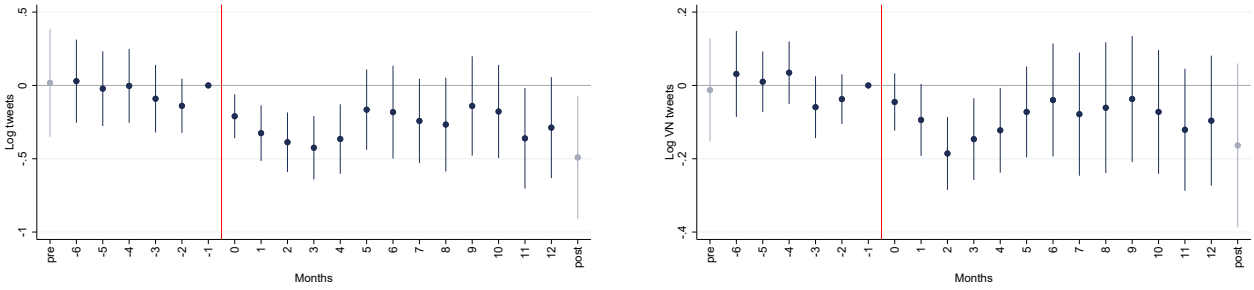
6 Tables and figures

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

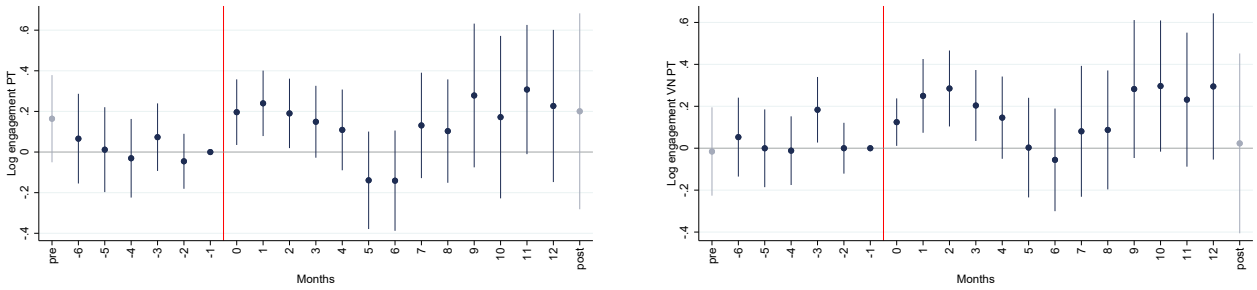


(a) All tweets, month × state FE

(b) Violence tweets, month × state FE

Note: All regressions include outlet fixed effects (FE) and FEs for state where an outlet is located × calendar month. Robust standard errors are clustered by outlet.

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

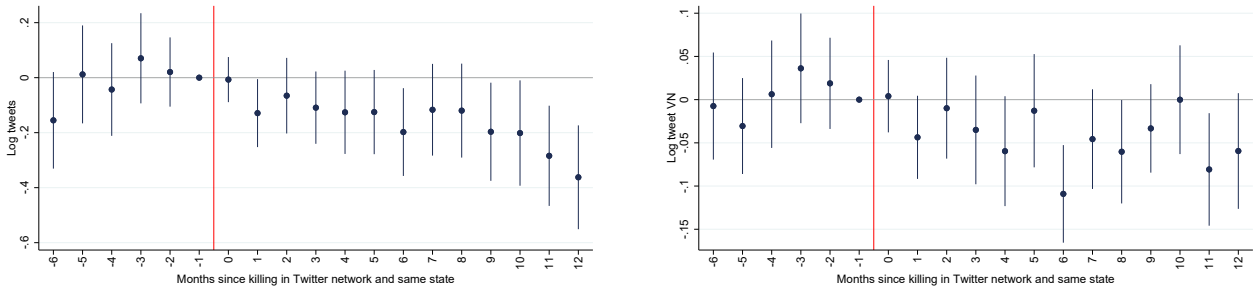


(a) All tweets, month × state FE

(b) Violence tweets, month × state FE

Note: Engagement is defined as likes and retweets normalized by the number of tweets for a given outlet-day combination. All regressions include outlet fixed effects (FEs) and FEs for the state where an outlet is located × the calendar month. Robust standard errors are clustered by outlet.

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

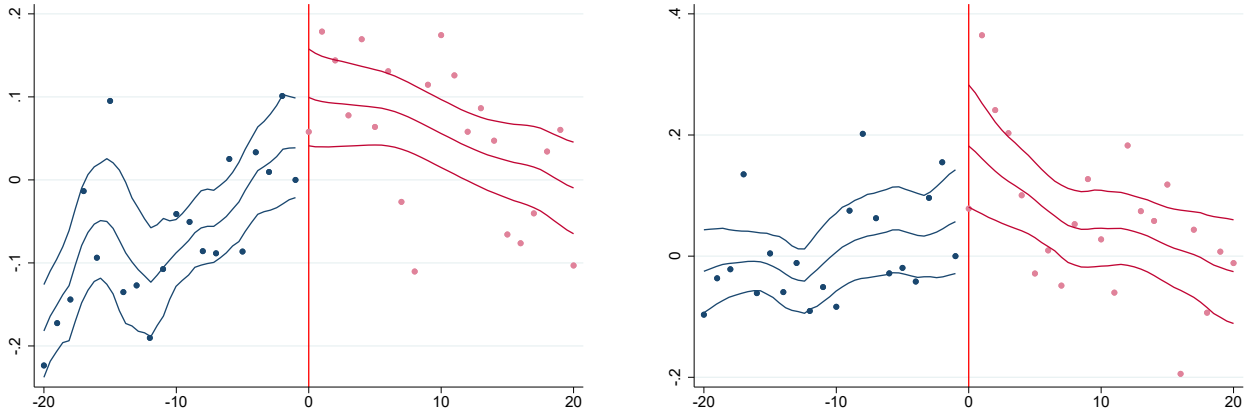


(a) All tweets

(b) Violence tweets

Note: Regressions compare Twitter activity of journalists in the same state that followed (on Twitter) murdered media workers with journalists in the same state that did not.

Figure 5: Google Trends search volume

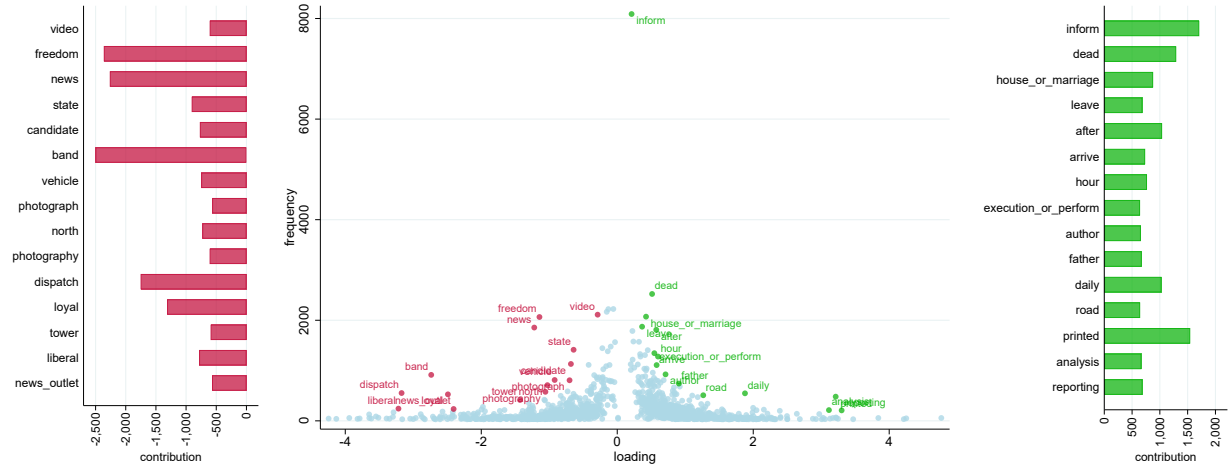


(a) Keyword: *murder*

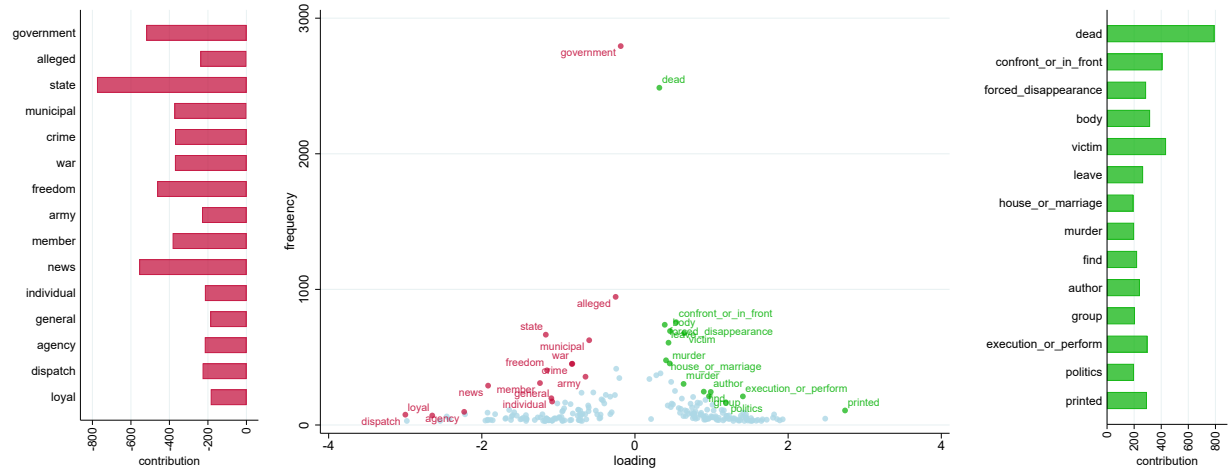
(b) Keyword: *journalist*

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

Figure 6: Terms that most predict timing of a tweet



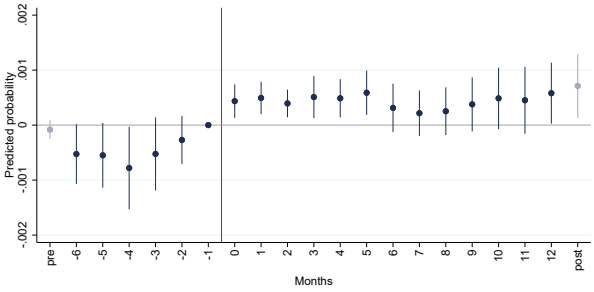
(a) All tweets



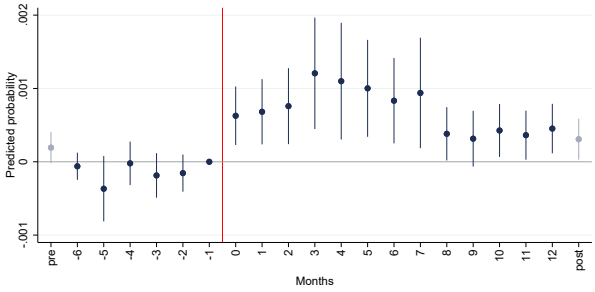
(b) Violence tweets

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

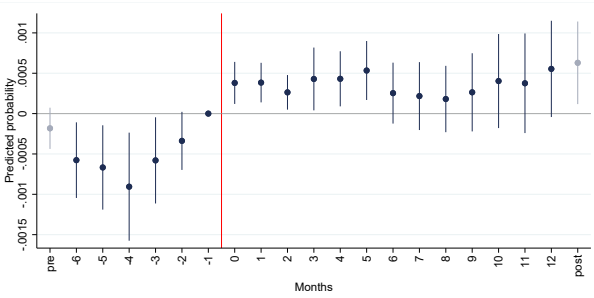
Figure 7: Tone of coverage among targeted outlets



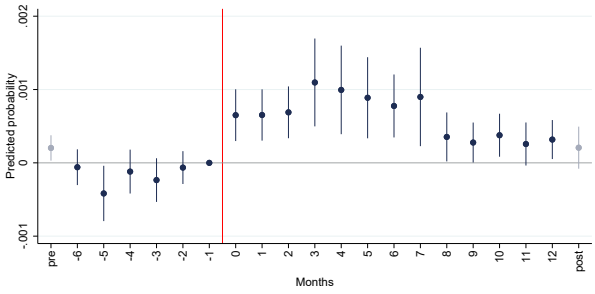
(a) All tweets, calendar-month FE



(b) Violence tweets, calendar-month FE



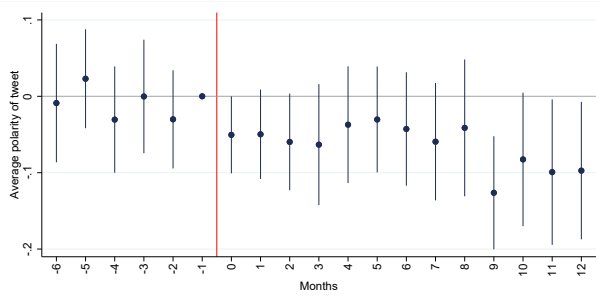
(c) All tweets, calendar-month FE $\times$ state FE



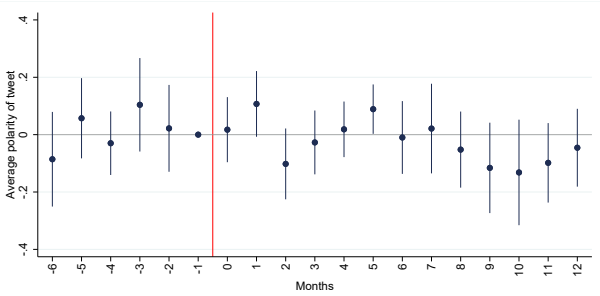
(d) Violence tweets, calendar-month FE $\times$ state FE

Note: We estimate a logit model in which 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 timing. 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 $7, \cdots, 12$ and *pre*, *post* are out of sample. Robust standard errors are clustered by outlet.

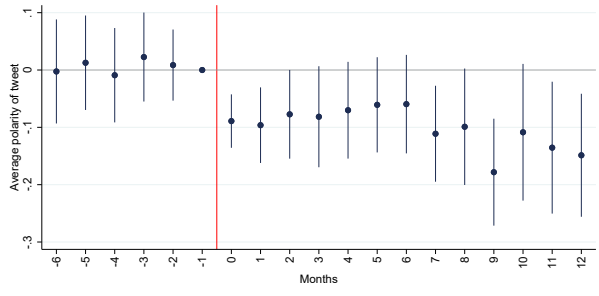
Figure 8: Changes in polarity following an attack



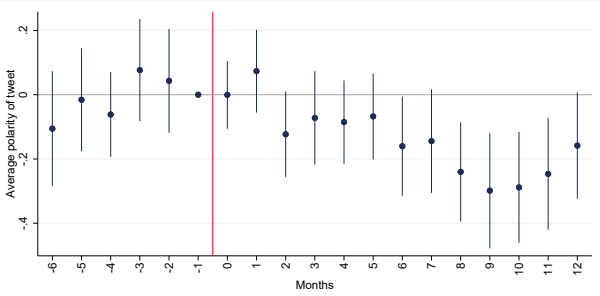
(a) All tweets, calendar-month FE



(b) Violence tweets, calendar-month FE



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



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

Note: We consider the set of terms with nonzero loading from the MNIR estimation that are also considered in the polarity dictionary of Brooke et al. (2009). Polarity is computed as the arithmetic average of polarity of tweets. We control for outlet fixed effects (FEs). Panels a and b control for calendar-month FEs, while panels c and d interact these FEs with indicators for the state where the outlet is located. Robust standard errors are 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 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 1 if individual reports being a journalist. Robust standard errors are clustered at the state level in parentheses. 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 are clustered at the state level in parentheses. Significance levels shown below *p<0.10, ** p<0.05, ***p<0.01.

Appendix

Data censoring

As discussed in section 3.3, we observe a large number of outlet-day combinations with twenty tweets, which might indicate censoring in the data. We address this by estimating an event-study specification that considers the number of combinations of outlets and days with data censoring as the left-hand-side variable (figure A6). We find 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 by a given outlet (larger values indicate that an outlet published tweets earlier in the day). As long as the distribution of tweets during the day remains unaffected by an attack, lower values (that is, tweets published closer to the end of the day) indicate a higher probability that more tweets were published on a particular day and higher values indicate a lower probability (Morales, 2020). Figure A7 shows a similar pattern to our main specification (figure 2): following the attack, there is an increase in average time elapsed (that is, fewer tweets) 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 CSE API to find articles for those outlets using the following list of keywords: *narco*, *ejecución* (execution), *fosas* (illegal grave), *cartel*. Our queries produced a similar set of results to what one would obtain from the following query: **site:outlet-webpage.com**

"keyword" range:$date_1$-$date_2$. Note that Google CSE is context aware, such that it returns matching articles even if they do not include the specific keywords, provided Google's proprietary algorithms determine that the articles are relevant to the query. The Google CSE API produces a maximum of 100 results for a given query. This limitation is problematic because it would lead us to underestimate the number of matches for queries with more than 100 hits. To address this issue we restrict our queries to thirty-day windows. Out of the 10,740 queries for outlet $\times$ thirty-day window $\times$ term, only 1 had 100 hits. Another empirical issue is that some newspaper websites were no longer online. In those cases, the 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 used to find the dates of articles worked 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 thirty-day query itself to assign dates, which reduces precision (for articles for which we can retrieve a date, Google's date range is accurate in more than 95 percent 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 data. The CSE correlates weakly with the EFIC database, which indicates issues with the CSE. To limit the issues, we consider as a dependent variable in our event-study estimates an indicator for outlet $\times$ thirty-day periods in which *any* articles were published.

Figure A9 depicts our estimates. We observe a fall of approximately 0.2 log points in coverage after an attack that peaks four months after the aggression. The coefficient for $t = 1$ is not statistically different from $t = -1$, but this could be because of initial reporting about the attack itself. Our estimates are thus in line with those using Twitter in section 4.2.

The MNIR model

Like many algorithms used in text analysis, MNIR seeks to predict the sentiment of a document based on the natural language. It's a supervised machine learning method, meaning that the researcher needs to provide a subsample of documents with their corresponding sentiment values.

Let y_i be the sentiment of a given document i, and x_i the corresponding "tokens" (in our settings, it would be a bag of words representation of the text). A common method to predict sentiment would be to estimate $y_i|x_i$. Since the dimension of x_i tends to be very large, methods to reduce the dimensionality are used, such as penalized regression. MNIR on the other hand uses an inverse regression (IR) approach, wherein the *inverse conditional distribution* for text given sentiment is used to obtain low-dimensional document scores that capture the relevant information from y_i.

Representation and sufficient reduction

Assume that there are n documents indexed by i and p "tokens" (words or bi-grams) indexed by j. The framework of interest is described by

$$x_i \sim MN(q_i, m_i) \text{ with } q_{ij} = \frac{e^{\eta_{ij}}}{\sum_{l=1}^{p} \eta_{il}},$$
$$\text{where } \eta_{ij} = \alpha_j + v_i' \phi_j$$

v_i is a random vector of y_i. In practice, we will take $v_i = y_i$. y_i is a $K - dimensional$ vector, although in our application it will be reduced to $K = 1$ where the feature indexed is a binary variable that captures whether a given tweet falls into the 180-day window following the murder of a journalist. $m_i = \sum_{j=1}^{p} x_{ij}$ is the total number of words in document i.

The idea behind MNIR is that one can use the projection of x on a lower dimension

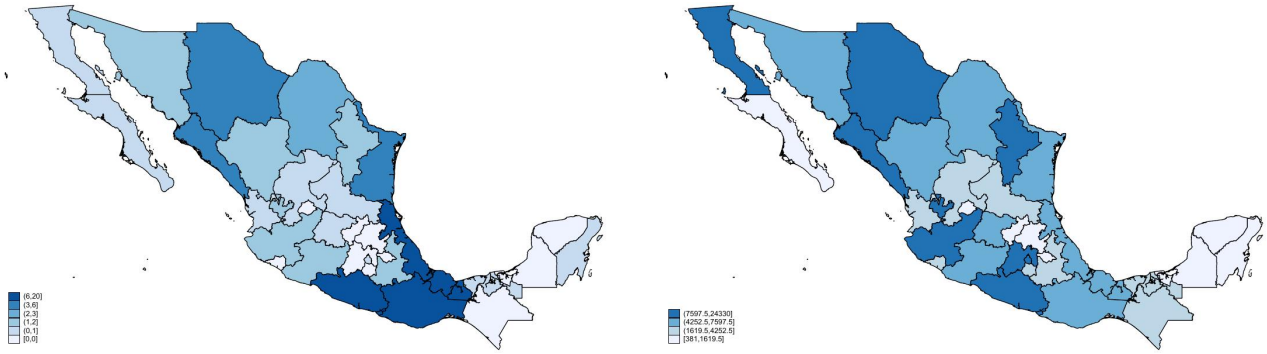
space through matrix multiplication with $\Phi = [\phi_1 \cdot \phi_p]'$ to build a "sufficient" projection. As follows from Taddy (2013) proposition 3.1: conditional on m_i, u_i: $y_i \perp x_i | v_i \rightarrow y_i \perp x_i | \Phi' x_i$. In other words, the projection $\Phi' x_i$ contains the same information on y_i as x.

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	.	.

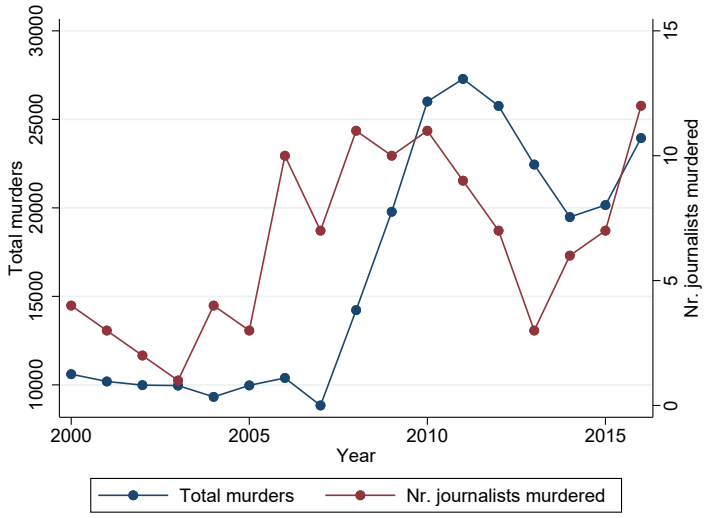
Figure A1: Homicides in the country (2009-17)



(a) Journalists (b) All homicides

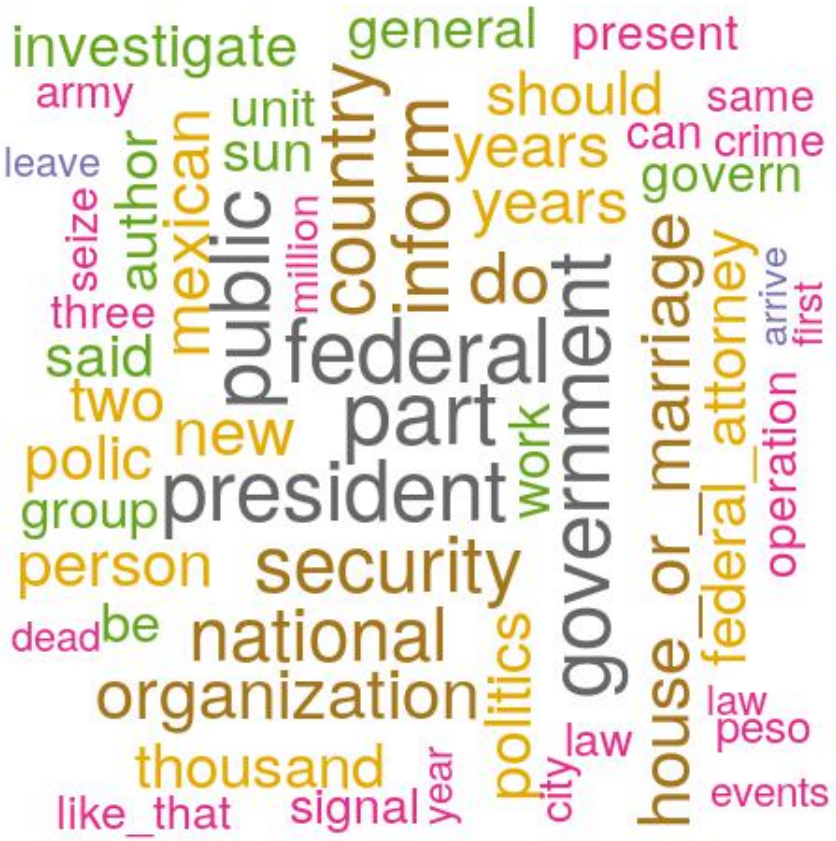
Note: Panel a depicts homicides of journalists, and panel b depicts total homicides in the thirty-two states of Mexico between 2009 and 2017.

Figure A2: Homicides



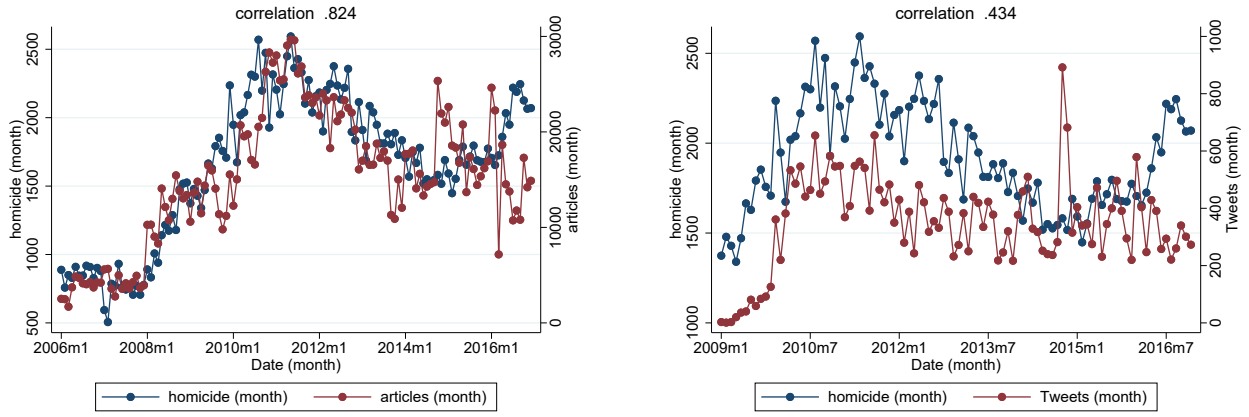
Note: This figure presents annual homicides in the country among the general population and the press.

Figure A3: Most common words (EFIC)



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

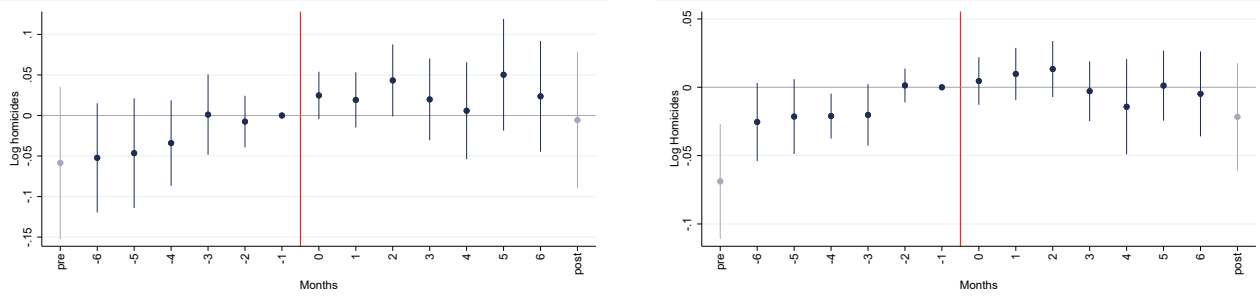
Figure A4: Monthly articles or tweets, and homicides in the country



(a) EFIC (b) Tweets about violence

Note: This figure depicts the time series of monthly number of articles published by the national press (EFIC, panel a) and the number of tweets published by the most important journalists in the country (panel b) against the number of total homicides. Panel b likely underestimates the true correlation because of data censoring, and coverage of the Massacre of Ayotzinapa (see section 3.3).

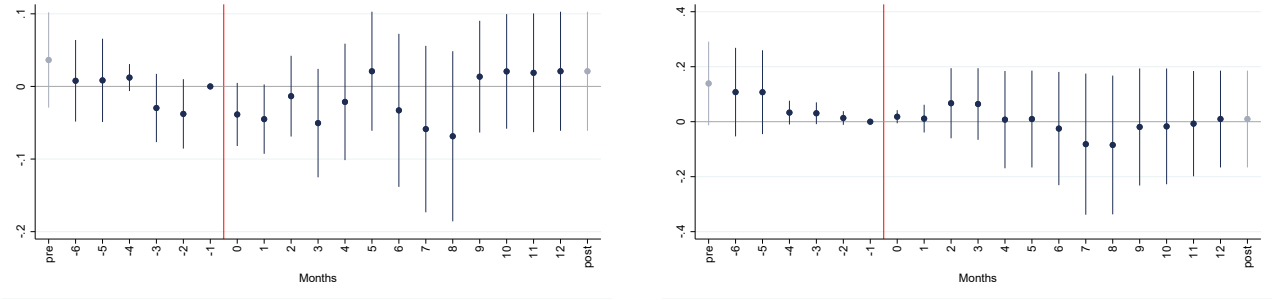
Figure A5: General-population homicides and timing of attacks against media workers



(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 are clustered by state or municipality.

Figure A6: Victimization and censoring among targeted outlets

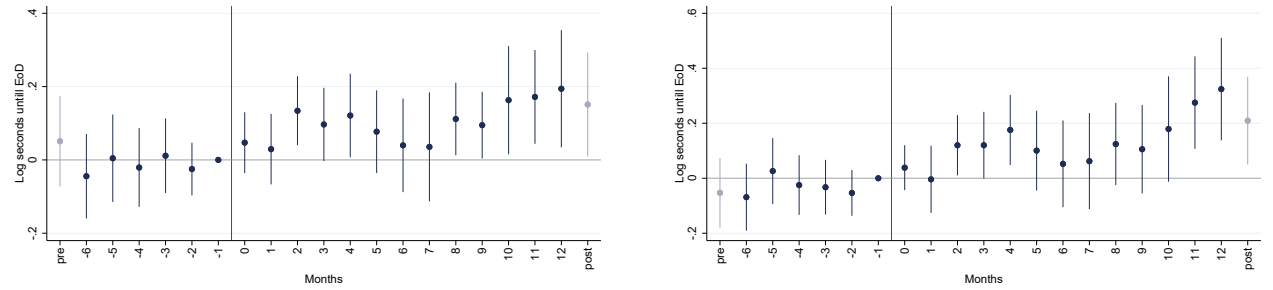


(a) Calendar month FE

(b) Calendar month × state FE

Note: The sample comprises outlets × days with twenty tweets, as this might indicate censoring. All specifications include outlet fixed effects (FEs). Robust standard errors are clustered by outlet.

Figure A7: Average timing of tweets of victimized outlets

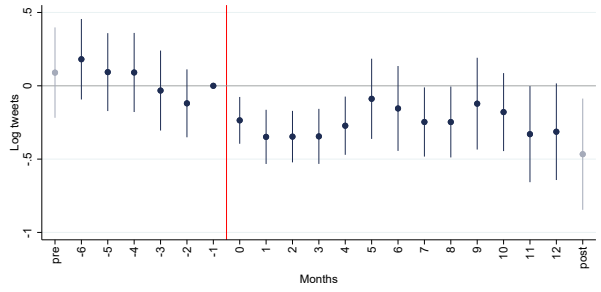


(a) Calendar month FE

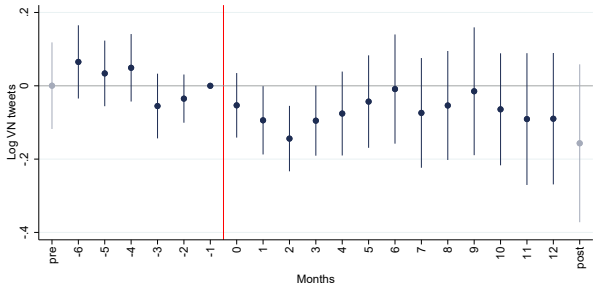
(b) Calendar month × state FE

Note: Dependent variable is the log of the average number 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 the distribution is independent of the attack, higher values indicate fewer tweets. All specifications include outlet fixed effects (FEs). Robust standard errors are clustered by outlet.

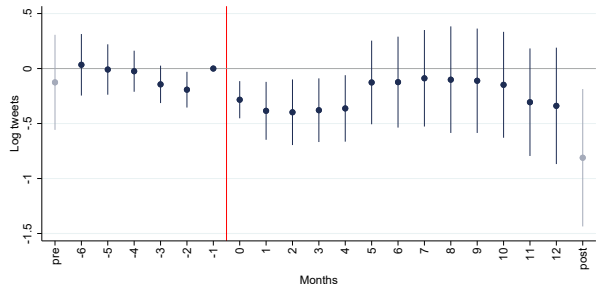
Figure A8: Direct effects of an attack on volume of coverage (alternative specifications)



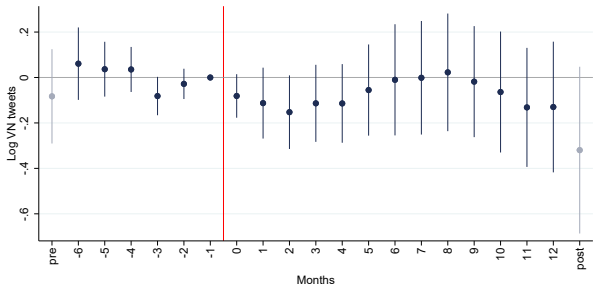
(a) All tweets, month FE



(b) Violence tweets, month FE



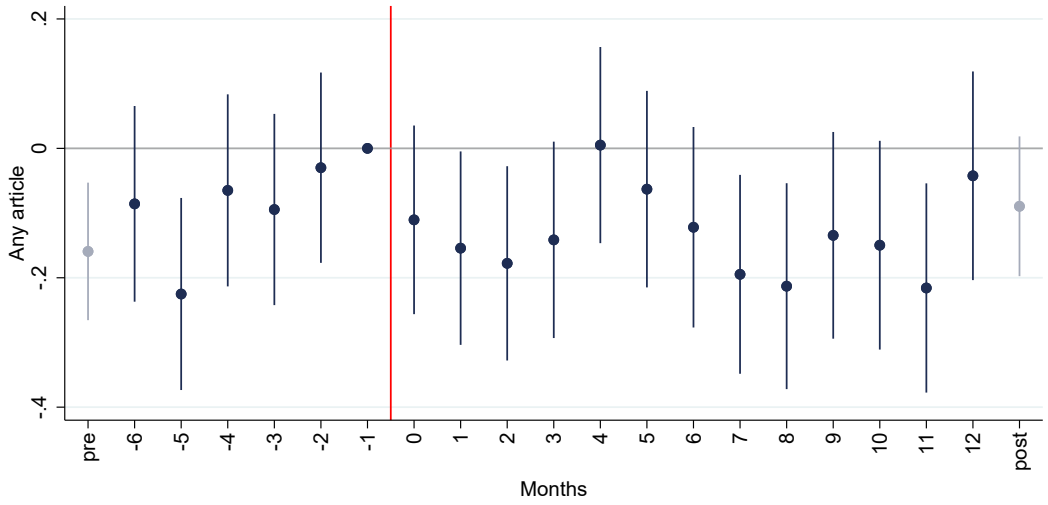
(c) All tweets, month × municipality FE



(d) Violence tweets, month × municipality FE

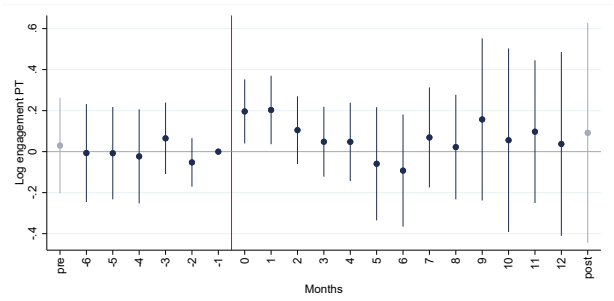
Note: All regressions include outlet fixed effects (FEs). Panels a and b control for calendar-month FEs. Panels c and d control for FEs for location municipality × calendar month. Robust standard errors are clustered by outlet.

Figure A9: Direct effects on coverage (Google Custom Search Engine)

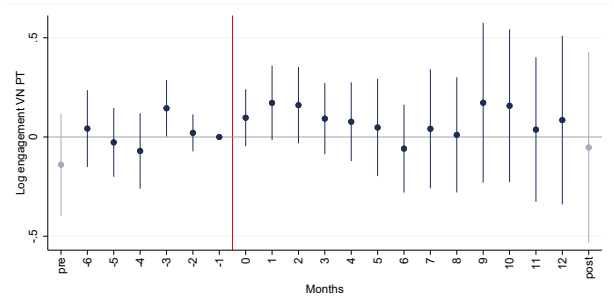


Note: This figure presents event-study estimates in which the dependent variable is an indicator for the publication of *any* articles. Outlet and thirty-day-period fixed effects are included. Robust standard errors are clustered by outlet.

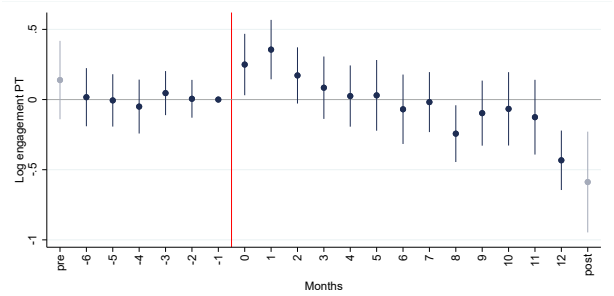
Figure A10: Direct effects of an attack on Twitter engagement with victimized outlets (alternative specifications)



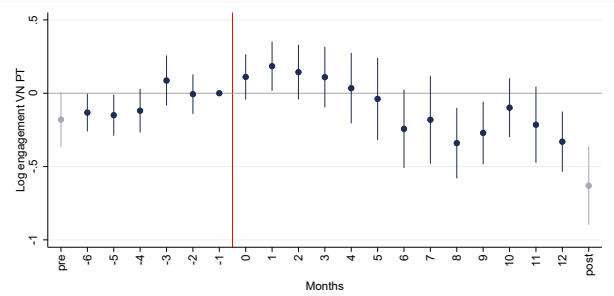
(a) All tweets, month FE



(b) Violence tweets, month FE



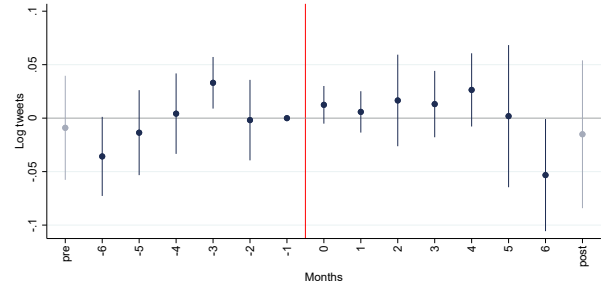
(c) All tweets, month × municipality FE



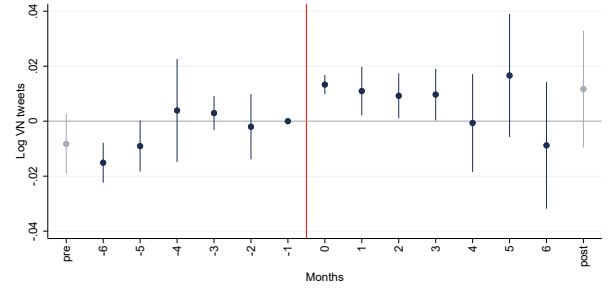
(d) Violence tweets, month × municipality FE

Note: Engagement is defined as likes and retweets normalized by the number of tweets for a given outlet-day combination. All regressions include outlet fixed effects (FEs). Panels a and b control for calendar-month FEs. Panels b and c control for FEs for location municipality × calendar month. Robust standard errors are clustered by outlet.

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



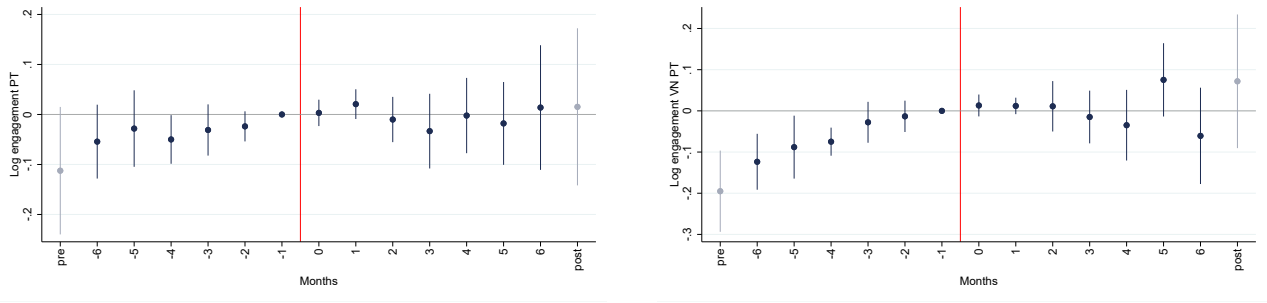
(a) All tweets



(b) Violence tweets

Note: All regressions include fixed effects for journalists × event and 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 are clustered by journalist.

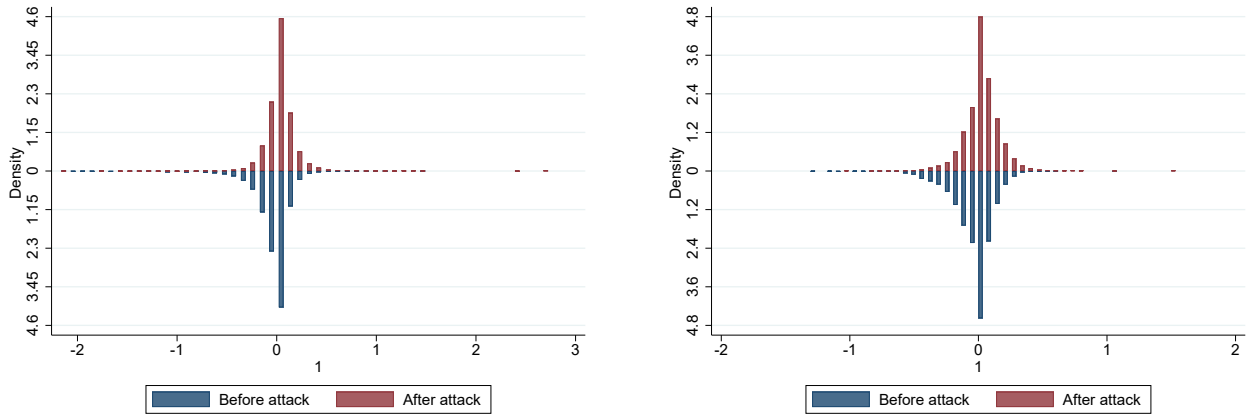
Figure A12: Indirect effects of an attack on engagement, top journalists



(a) All tweets (b) Violence tweets

Note: Engagement is defined as likes and retweets normalized by the total number of tweets per journalist-day. All regressions include fixed effects for journalists × event and 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 are clustered by journalist.

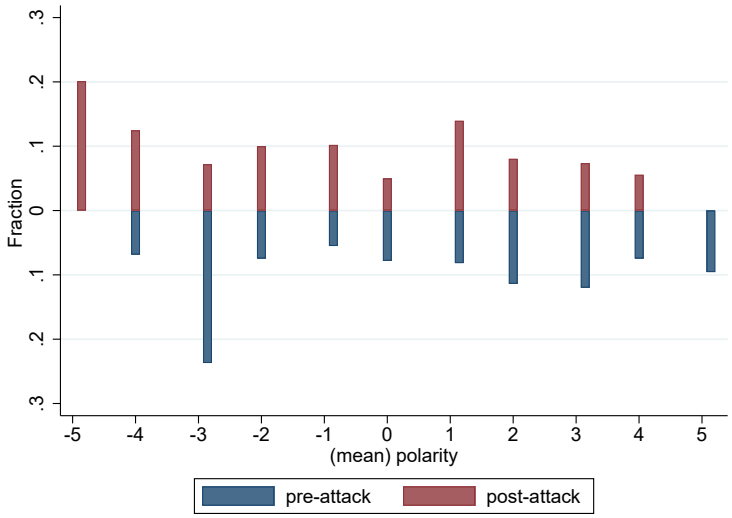
Figure A13: Distribution of sufficient reduction by timing of attack



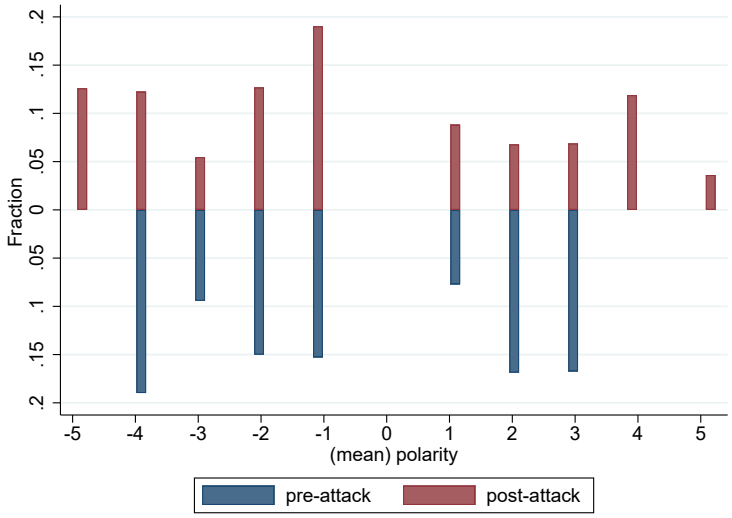
(a) All tweets (b) Violence tweets

Note: Depicted here 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 between 180 days before and 180 days after the homicide of a journalist.

Figure A14: Distribution of terms by polarity



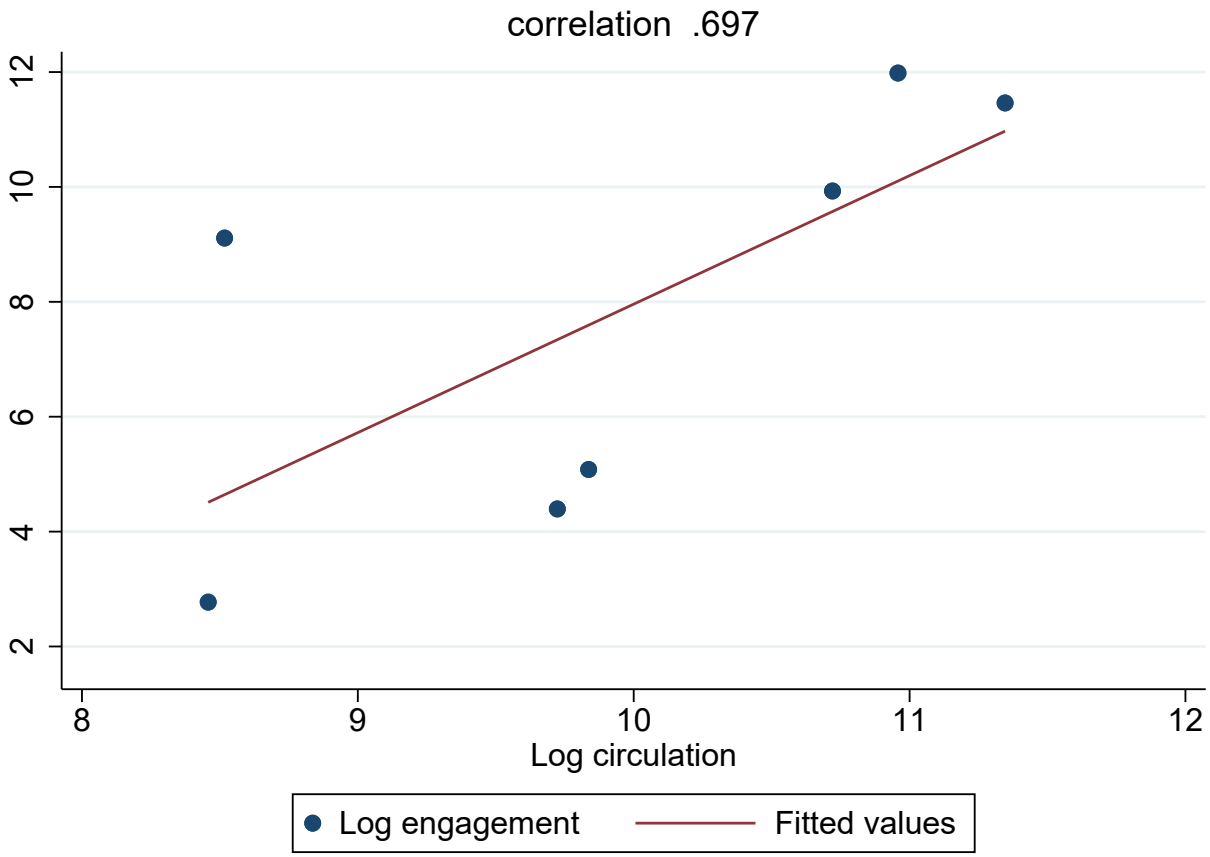
(a) All tweets



(b) Violence tweets

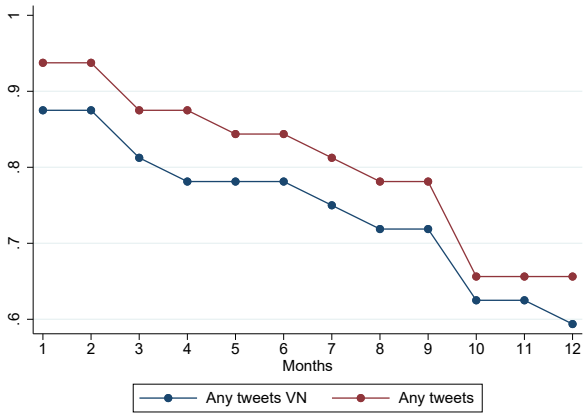
Note: These figures show the distribution of polarity in tweets from targeted outlets. Considered are terms with nonzero coefficient in the MNIR regression between 180 days before and 180 days after an attack. Polarity is computed from the polarity dictionary by **?**.

Figure A15: Circulation and Twitter engagement

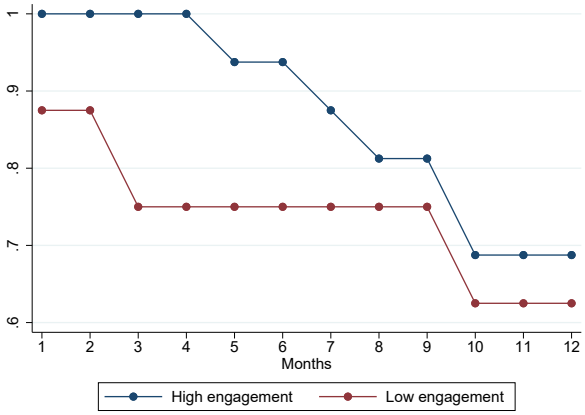


Note: Engagement is defined as likes and retweets. We calculate engagement using tweets six months before an attack on a news outlet, as the event might have an independent effect on engagement. Circulation figures come from the State Secretariat's census on the media.

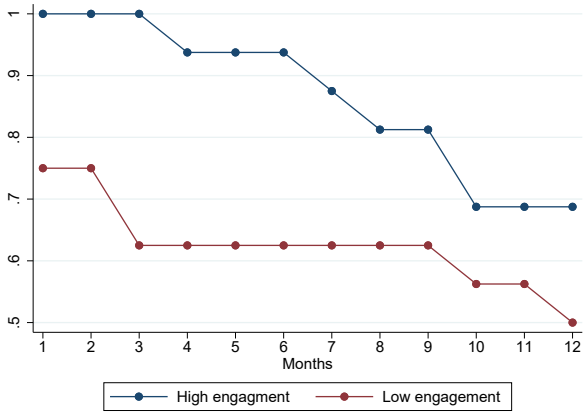
Figure A16: Survival rates of victimized outlets after an attack



(a) All and violence tweets



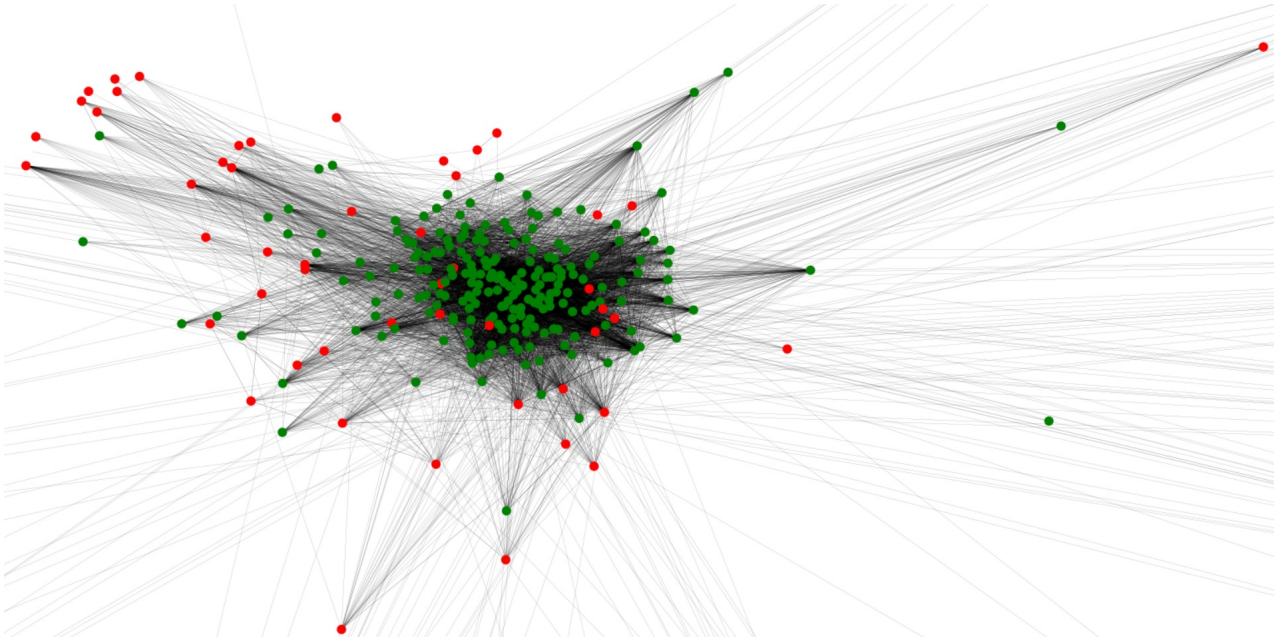
(b) All tweets



(c) Violence tweets

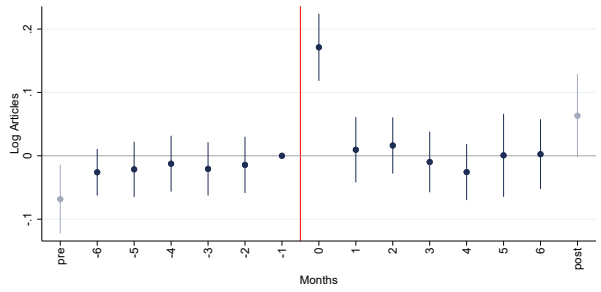
Note: This figure presents the probability that a targeted outlet tweets at least once between month j after an attack and month 12, with $j \in [1, \cdots, 12]$. Only outlets with at least one tweet in the thirty days preceding an attack are considered. Tweets on the day of the attack are ignored, as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

Figure A17: Twitter network of selected accounts

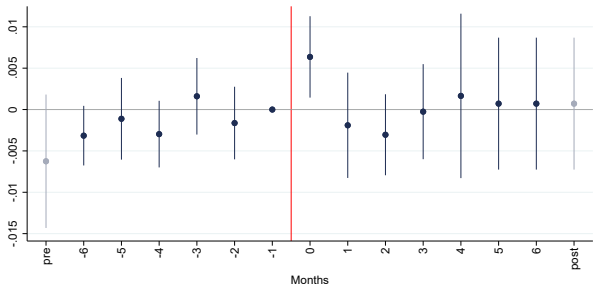


Note: The figure shows a partial Twitter network for the accounts in our dataset. Red nodes represent victimized outlets, and green nodes represent journalists. An edge is drawn between two nodes if either of the accounts follows the other.

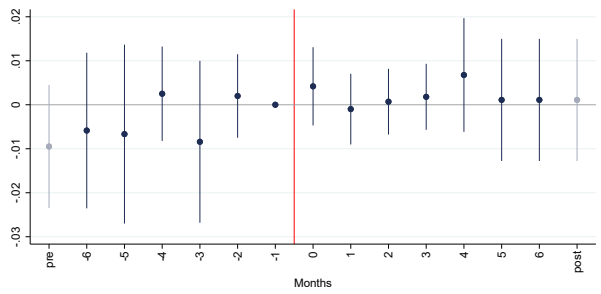
Figure A18: Mentions of municipalities in the national press



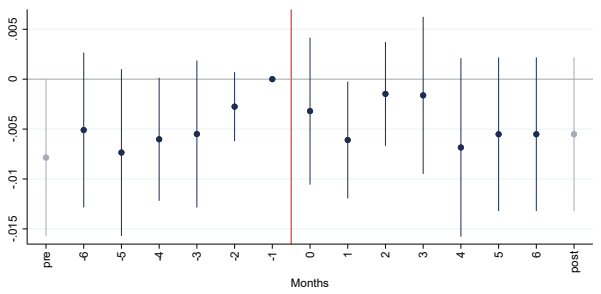
(a) Municipality



(b) Municipality + *hitman*



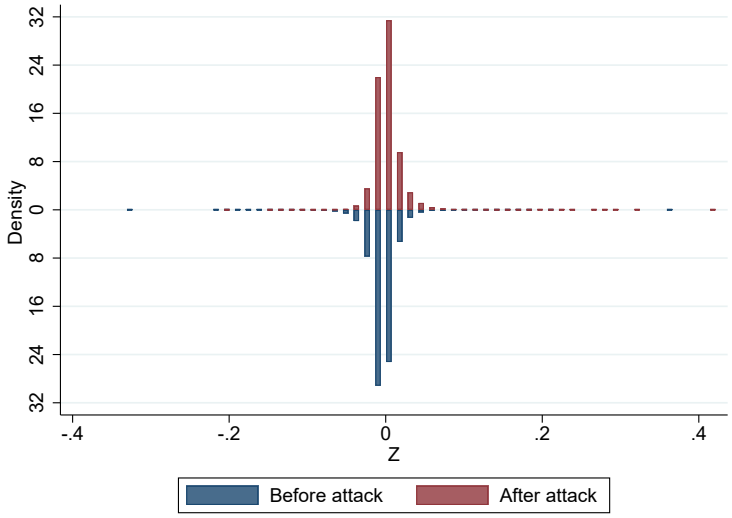
(c) Municipality + *Sinaloa Cártel*



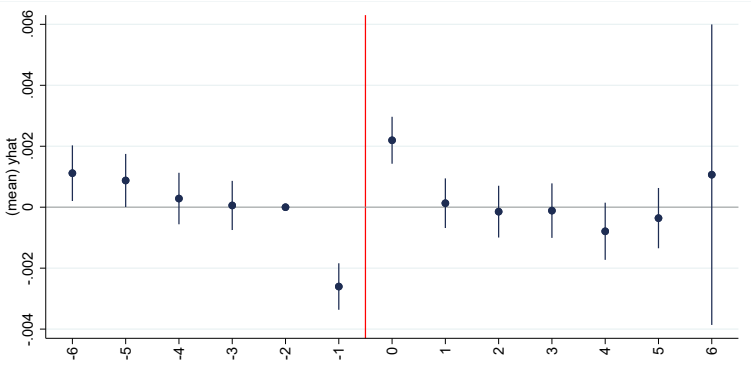
(d) Municipality + *Jalisco Cártel*

Note: This event study considers mentions of municipalities in the national press before and after a journalist was killed there. Panels 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 the United States.

Figure A19: 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 between thirty days before and thirty days after an attack. This figure presents the distribution of the sufficient reduction (SR) statistic (panel a). Panel b shows the predicted probability of post-attack timing from a linear probability model in which we regress the post-attack status on the SR.