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

October 27, 2022

Abstract

This paper studies how the murder of journalists affects news media in Mexico. We use data collected from a comprehensive archive of large news organizations and over nineteen million tweets published by journalists and media outlets between 2010 and 2020. Using a series of event-study analyses, we find large reductions in reporting activity among victimized outlets following journalists' killings. This reduction is consistent with self-censorship but the "effectiveness" of attacks as a censorship tool wanes over time. We also find subtle but persistent changes in the tone of coverage. Using census data we show that states with more violence against media saw reductions in the number of active journalists and changes in their demographics. The empirical patterns suggest that *selection* out of the profession is an important driver of the dynamics in this context. We formalize these ideas in a theoretical framework of news reporting in violent settings.

Word count: 9,326

1 Introduction

Freedom of the press is one of the fundamental pillars of democracy, playing an essential role in informing voters and in keeping governments and politicians accountable. However, in the last decade more than 550 journalists were killed around the world, threatening this freedom and weakening democratic values (Carey and Gohdes, 2021). This paper studies the relationship between targeted violence against the media and news reporting in Mexico, one of the most dangerous countries for journalists.

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[1] could have mixed effects on media activities. On one hand, violence may deter the press from covering certain events through fear and self-censorship. A Mexican journalist working along the US border reported:

> "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 induce journalists' *backlash*, as suggested by another reporter:

> "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)

These potentially contrasting responses determine the extent to which violence against journalists constitutes an effective censorship tool. However, our understanding of the overall effects of violence against journalists on media activity remains limited. We contribute to this understanding by studying the effects of journalists' killings in Mexico. We use a new Twitter dataset of over 19 million tweets published by Mexican media accounts

[1] We use the terms "media worker", "journalist" and "reporter" interchangeably.

over the last decade, and data from a private database (Eficiencia Informativa) with over thirty-five million news articles from Mexican outlets. In addition, we use data from the 2010 and 2015 Mexican census. Finally, we combine these sources with data on all reported killings of members of the press in Mexico between 2009 and 2020.

Our main analysis uses 70 Twitter accounts of media outlets that had one or more affiliated journalists murdered. Using a difference-in-differences specification with staggered treatment and heterogeneity in effects across time, we find that journalists' killings led to substantial reductions in media activity. In the 6-months after the killings, victimized media outlets published significantly less tweets. However, these effects have become smaller over time and are not statistically significant by the end of our period of study. This evidence suggests that outlets which remain active in Mexico have become more resilient and less responsive to attacks over time.

In our main empirical exercise we exploit the precise timing of journalists' killings in a series of event-study exercises. Using Callaway and Sant'Anna (2021)'s recent contributions to the difference-in-differences methodological literature, we measure short and medium-run changes in news activity following the violent incidents. This analysis also reveals sustained reductions, of around 25 percent, in victimized outlets' Twitter activity following the murder of an affiliated worker.

We then use a modern supervised machine-learning algorithm and the text of the accounts' tweets, to quantify changes in *tone* among victimized outlets following an attack. 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. These results also provide some evidence of *backlash*, suggesting that outlets that remain active despite the violence become more critical, but the effects are subtle and relatively small in magnitude. We also show that following the killings, grammatical errors become more common.

Using data from the Mexican 2010 and 2015 census we then study long-term changes in the demographics of journalists. We identify active journalists (based on occupational codes) in the thirty-two states of the country. We implement a long difference-in-differences design to estimate how violence against media workers affects the journalism occupation overall. We first show that states with more homicides have fewer total active journalists in 2015 relative to 2010. Violence against journalists is also associated with significant changes to the demographic characteristics of journalists, with reporters in more violent states being less likely to be married or have children in 2015 relative to 2010. These patterns reveal that there is significant selection in who chooses to exit and who chooses to remain in the profession over time.

Finally, we present a formal model of reporting and violence that helps frame our results and put them in a broader context. Our model captures two important ideas of the environment in question. First, consistent with qualitative evidence, there are different *types* of journalists who are more or less sensitive to violence: some would feel dissuaded from reporting due to fear, while others would feel more commited to their work. Second, the empirical patterns we document suggest that *selection* out of the profession is an important driver of the dynamics in this context. We formalize these ideas through a simple framework in which journalists face both contextual and targeted violence, and make decisions on whether to report news or not. As contextual violence increases, less resilient journalists leave the profession, in turn this makes targeted violence less effective at censoring the press. The model captures the main features of our empirical results.

Our work is related to the political economy literature on the effects of censorship and the determinants of media coverage. The importance of the media for political processes has been widely established.[2] Several papers have studied different 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 et al. (2020) and

[2] See for instance, McMillan and Zoido (2004); Dellavigna and Kaplan (2007); Snyder and Stroemberg (2010); Adena et al. (2015, 2021); Foos and Bischof (2022)

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 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. Our paper documents that *violence* can also be a significant determinant of media content. In addition, our use of social media data to document patterns in news media activity also contributes to an emerging strand of the literature on the political economy of media that adopts online data sources (which includes Cagé, Hervé and Mazoyer, 2022; Cagé, Hervé and Viaud, 2020; Hatte, Madinier and Zhuravskaya, 2021; Ershov and Morales, 2021, among others).

Most previous studies on press repression and media activity have 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 limited (Pan and Siegel, 2019). In Venezuela, public anti-media messages by government officials are a significant predictor of physical attacks against the media (Mazzaro, 2021). Most violence against journalists, however, occurs in democracies (Carey and Gohdes, 2021), where cultural values regarding freedom of speech could potentially mitigate the perverse effects of this violence on press activity. Our article presents a systematic study on the effects of violence against journalists in the context of a recognized electoral democracy.

Perhaps not surprisingly, several papers study repression of journalists in the context of 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. 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. Finally, Dorff, Henry and Ley (2022) studies the effects of violence against journalists on the media activity of a particularly renowned national outlet, showing how media risk is associated with less specificity in its news reports. We contribute to this literature by presenting a systematic study of killings of journalists in Mexico over the past decade and how these affected the activity of targeted outlets. More generally, we study how violence against journalists has changed the demographics of the profession.

The paper is organized as follows. Section 2 discusses the political and media context of Mexico. Section 3 presents our data sources and some descriptive statistics. Section 4 outlines our empirical strategies and presents the results of the analysis, documenting how the killings of journalists affected news media, as well as robustness checks and empirical extensions. Section 5 presents a theoretical framework on media workers in contexts of violence. Section 6 concludes.

2 Context

Between 2006 and 2020 the Committee to Protect Journalists (CPJ) and Article 19 documented the murder of 143 media workers in Mexico, making it one of the deadliest countries for journalists worldwide. These homicides are highly targeted operations and in a majority of cases victim journalists covered *crime*. According to CPJ, at least 15 journalists were kidnapped and tortured before being killed, and at least 39 of them received threats as well (*ibid*).

Criminal organizations often try to influence news reporting and in many cases these murders were punishment for individuals or outlets being "out of line". Valdez (2016) documents:

> "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." [p. 42] [3]

Figure 1 presents the proportion of victims that reported on various press topics (based on data from the CPJ): 76 percent covered crime or police, 45 percent politics, and 18 percent corruption (these figures do not add up to 100 percent because some journalists covered multiple subjects).

This high degree of violence is a recent phenomenon and is linked to the start of the Mexican War on Drugs. Figure A1 shows how the targeting of media workers increased steeply in 2006, followed two years later by an increase in homicides among the general population. 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 hard-line military strategy against drug-trafficking organizations was underway; the Federation splintered into two bands fighting for control of profitable drug-trafficking routes (Hernández, 2012). These events coincide with the increase in killings of media workers and members of the public that persists to this day.

In figure A2, panel *a*, we observe how the states of Guerrero (south-west, in dark blue), Veracruz (east, in dark blue) and Oaxaca (south, in dark blue) report the largest number of journalists killed. These states alone account for 42 percent of all journalists killed, but only for 13.2 percent of the country's population.[4] Guerrero is the state where most poppy is eradicated, and number 5 in terms of marijuana; Veracruz is 17 and 16, respectively; and

[3] Own translation. Valdez himself was also killed following the publication of his book.

[4] See INEGI (consulted May, 2022). https://www.inegi.org.mx/app/tabulados/default.html?nc=mdemo02

Oaxaca 6 and 7[5]. Veracruz alone accounts for 18.9 percent of all deaths of journalists. The state is the sixth in terms of seizures of cocaine. Every other state with larger seizures is either a border state with the US or a major touristic destination. This likely means it is an important entry point of the illegal narcotic (*ibid.*).

In addition to being states with important domestic production of narcotics and an entry point for South American cocaine, Veracruz and Guerrero were contested by more than one cartel as of 2015.[6] Holland and Rios (2017) find that inter-cartel rivalry, more so than the mere presence of cartels, increases the likelihood of an attack on the press in the country.

Importantly, the pattern of violence against the press stands in contrast with the overall level of homicides in the country, which roughly tracks states with significant production of narcotics or states that are themselves part of drug-trafficking supply routes to the United States.

One reason behind this high level of victimization is the low number of cases that are solved by the police (Carey and Gohdes, 2021). 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:[7]

> 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.

[5] These figures are reported by the Mexican Army for the years 2000-2012.

[6] Source: Drug Enforcement Agency. *(U) Mexico: Updated Assessment of the Major Drug Trafficking Organizations' Areas of Dominant Control*. DEA-DCT-DIR-064-15

[7] 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 Data

3.1 Attacks on journalists

We obtain data on attacks on journalists from Article 19 and CPJ, two leading NGOs that advocate for journalists. These organizations keep track of murders of media workers in Mexico, along with the workers' affiliations. Out of 143 victims, 11 were classified as freelancers and hence are not matched to any outlet. The rest are assigned to 155 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.

3.2 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.[8] The second set of usernames corresponds to 116 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 2020. Our final datasets comprise 9.3 million tweets published by Mexican journalists and 9.7 million tweets published by victimized outlets.

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 10.4 percent of tweets in the journalists dataset, and 13.3 percent in the outlets dataset.[9] Figure A3 panel

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

[9] The set of words is the following: *cartel*, *narco*, violence, homicide, death, body, threat, justice, alleged,

b shows the number of tweets with violent content by the top journalists and number of homicides in the country. This measure 0.12 correlation is likely underestimated because of an almost five-fold increase in tweets in December 2014, which was likely caused by coverage of the massacre of Ayotzinapa.[10]

3.3 Other data

We gathered data on daily-circulation of thirty-five victimized outlets from the State Secretariat's National Census of Printed Media[11] (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.[12] Finally, we access data on the 2010 and 2015 Mexican censuses through IPUMS International.[13] The analyzed sample is restricted to journalists and occupational categories that are similar to journalists'; this includes accountants, researchers, psychologists, 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.

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.

[10] See for example, this article from the New York Times (accessed July 29, 2022)

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

[12] The Centro de Investigación y Docencia Económicas (CIDE) Drug Policy Program (PPD) maintains a database with homicides specifically attributed to drug-trafficking organizations compiled by a government panel, however, as the Mexican government ceased to track murders of this kind in 2011, we decided against using it.

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

3.4 Data discussion

We document how violence against media workers affects 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. Twitter data has the advantage of being widely available and high-frequency in nature. It has also been shown that online media activity is correlated with traditional activity (Cagé, Hervé and Viaud, 2020; Ershov and Morales, 2021) and that online activity can affect traditional mainstream activity (Cagé, Hervé and Mazoyer, 2022). Figure A12 shows that in our context, social media engagement is also correlated with newspaper circulation (for a small set of outlets for which this data is available). Furthermore, in section A.4 we show also that our results are similar if we consider a smaller set of articles that we retrieved using Google Custom Search Engine (CSE) API. Therefore, we view the results using Twitter as a good proxy for the effect on newspaper activity more generally.

4 Empirical Analysis

4.1 Baseline analysis

We begin our analysis of the effect of journalists' killings on media activity through a simple difference-in-differences model. We exploit the precise timing of killings and the panel structure of the data to estimate these effects. However, a recent literature in economics has raised concerns about difference-in-differences estimations with staggered treatment and how to interpret such estimates (De Chaisemartin and d'Haultfoeuille, 2020; Sun and Abraham, 2020; Sant'Anna and Zhao, 2020; Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021). We present a modified estimator for our baseline analysis to address some of these concerns.

One worry in our context is that the killings may affect not just the levels of activity, but also the trends. That is, that the activity of a victimized newspaper grows faster or decreases over time relative to non-victimized outlets. For this reason, we focus on relatively short-run outcomes: the six-months after the killing. Our *Post* indicators are equal to one only for the six months after the killing, and not for the remaining time in the sample. A second important concern is that there may be heterogeneity across treated cohorts, as the new econometric literature has shown that difference-in-difference estimates from two-way fixed effects models, which are weighted averages of all possible 2x2 diff-in-diff comparisons, can be sensitive to this form of heterogeneity. We include a linear time trend interacted with our main *Post* indicator to account for treatment heterogeneity across time. Finally, we include specifications which remove from the sample treated units after they are treated. This restriction ensures that we form a 'rolling control group' such that only non-victimized outlets enter the sample in our comparisons (see for example Cengiz et al., 2019).

Note that these are not perfect remedies and in the following section we present an alternative model. However, this difference-in-differences model provides a good baseline for the estimated effects and is intuitively easy to understand. Specifically, our main estimating equation for this analysis is a two-way fixed effects model of the following form:

$$\begin{aligned} y_{ost} &= \beta_1 * Post_{ost} + \beta_2 * Post_{ost} * Time_t \\ &\quad + \alpha_1 * Window_{ost} + \alpha_2 * Window_{ost} * Time_t + \delta_o + \delta_t + \varepsilon_{ost} \end{aligned}$$

Our outcome of interest, y_{ost}, is the number of tweets published by outlet o, located in state s, at time t (month x year). The variable $Time_t$ is a linear time trend that measures years since 2009 (the start of our sample). Our specifications include outlet and time fixed effects (δ). Our estimates therefore measure within-outlet changes in publishing activity. As mentioned above, $Post_{ost}$ is an indicator equal to 1 in the six months after the killing of an affiliated media worker. Our parameters of interest, β_1 and β_2, respectively capture the effect of a killing in the 6-months after the event, and the differential change in this effect as time passes.

In our preferred specifications, we also include an indicator variable $Window_{ost}$ equal to 1 for the one-year window around the event (6 months pre and 6 months post), as well as its interaction with the linear time trend. Including this variable *normalizes* the estimated effect measured by β relative to the six months before the event. That is, β measures the change in outlet activity relative to the time just prior to the killing. We also exclude this event-window indicator in specifications tests. Finally, we also include a state by year fixed effect in some specifications, such that the changes in activity are measured against other outlets in the same state, and not just all other outlets. Standard errors are clustered at the media outlet level.

The results from our analysis are presented in Table 1. The estimates indicate that journalists' killings have an effective censorship effect, as outlets reduce their activity in the months following a murder. The estimates β_1 suggests that victimized outlets publish 61 percent less tweets in the six months after an affiliated journalist is killed, relative to the six months before the killing. However, this estimate is that for the earliest year of our sample. The effect becomes smaller over time, as indicated by the positive and statistically significant positive coefficient measured by β_2. In fact, our estimates suggest that the effects dissipate completely by around 2017.

The estimates are somewhat sensitive depending on the specification but always statistically significant. In column 3, our initial estimates are larger (89 percent reduction) when removing treated units from the control group. Column 4, which does not include the event-window, reveal that relative to the entire pre-attack period the drop in activity is even larger (112 percent). The estimates in columns 6-10 include only tweets classified as those about violence or crime. The magnitudes are smaller for these estimates, but overall, the results are qualitatively similar.

4.2 Average Dynamic Treatment Effects

Our main estimates of the effects of killings on media activity exploit once again the timing of these events, but we do so in a flexible event-study framework that allows us to evaluate the usual parallel trends assumptions and the dynamic nature of the effects following the killings. In particular, we follow Callaway and Sant'Anna (2021) to estimate dynamic ATTs of the general form:

$$\theta = \frac{\sum w_{g,t} \times ATT(g,t)}{\sum w_{g,t}}$$

where:

$$ATT(g,t) = \Big[EY(g)_t - EY(NT)_t\Big] - \Big[EY(g)_{g-1} - EY(NT)_{g-1}\Big]$$

This is a weighted treatment effect estimate where $ATT(g,t)$ corresponds to a 'cohort' of outlets affected on the same *month-year*. Importantly in our case, given the relatively high-frequency nature of our data, most of these are cohorts of one single outlet. The $ATT(g,t)$ estimate compares the outcomes of these outlets to those of the not *yet* treated outlets (NT) at that particular point in time.

Figure 2 shows our estimates. Within two months of the killings, the volume of tweets decreases by around 25 percent for the full sample, and 5 percent for violence/crime tweets. Publishing falls to its lowest point around three to four months after the events, which could be partially explained by initial reporting about the attack offsetting a reduction in activity (see below on short run effects). We do not observe an indication of significant time trends before the killings. However, we do observe slightly higher publishing activity in the months just prior to the killing. One potential explanation for this 'blip' is that it is precisely this activity which the perpetrators were trying to dissuade.

In online appendix Figure A5 we show that the estimated dynamic effects are very similar if we include state x time fixed effects, a more strict specification which accounts for other events co-occurring in the same state as the killing. However, we lose some statistical power from this exercise due to the fact that in some states there is only one killing.

Overall, the results from the event study confirm our baseline estimates. We find that violence against journalists is an effective censorship tool, and that affected media outlets reduce their publishing activity in the months after an affiliated journalist is killed.

4.3 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 substantial 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 from large national outlets which are generally not targeted by criminal organizations.

Figure A7 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[14], 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 17 percent increase in mentions of the municipality. The effect subsides after the first month, and we find no evidence of pre-trends before the event.[15]

In a similar fashion, we look at mentions of pairs of municipalities and keywords, such as "hitman," "Sinaloa Cártel" and "Jalisco Cártel."[16] 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

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

[15] That the increase in attention from the attacks tends to be short-lived is consistent with evidence from other contexts (Morales, 2021; Krakowski, Morales and Sandu, 2022).

[16] 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.

organizations (panels *c* and *d*).

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 machine learning algorithm using a 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 predicting the timing of a tweet is evidence that language changed as a result of an attack. We characterize the change in language by identifying the words that contribute the most to determining the timing of a tweet. 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 Twitter text, a medium that follows language conventions loosely, we perform an exhaustive text clean-up procedure. Section A.5 formally presents the model and we discuss the text pipeline in section A.7.

We train the MNIR model using the tweets published by targeted outlets. 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 would not be informative. Specifically, we filter out tweets that either mention the words "journalist" and "murder" or the name of a murdered media worker. The MNIR algorithm assigns zero values to two thirds of words in the corpus, which indicates that they are not useful for predicting.

Our first step in this empirical analysis is to assess the ability of the model to correctly predict the timing of a tweet. To do so we rely on the *Sufficient Reduction* (SR, denoted by

Z_{oi}) statistic.[17] Our outcome variable in this setup is equal to one for a tweet post-attack and zero otherwise. Thus, terms with positive Z values are associated with the post-attack time-frame and vice versa. If outlets used completely different sets of words before and after an attack, there would be two non-overlapping distributions of SR. Figure A10 presents the cumulative distribution of SR for tweets before and after an attack took place. We observe considerable overlap between the two, with post-attack tweets stochastically dominating. Any changes in language that are occurring are likely small and subtle, and are probably concentrated among violence tweets, as there is less overlap among these.

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 4 presents the distribution of loadings and frequencies for the full sample of tweets and the subsample of violence tweets. The top 15 terms that most contribute to identifying the timing of tweet are highlighted, where contribution is defined as the absolute value of product of loading and frequency.

Panel *a* shows that words most useful for predicting pre-attack are mostly unrelated to violent events. For instance, "video," "Mexican," and "debate", among others. In contrast, we find terms that likely reference violent events after an attack, including "murder" and "investigate". 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* addresses this by only considering the sample of tweets with violent content. Terms associated with the pre-attack time-frame are often used to report on law-enforcement operations: "government", "police", "law-enforcement". Other terms are

[17] Taddy (2013) describes conditions under which the SR performs as a summary of the available information pertaining to the dependent variable (see Section A.5).

more ambiguous, such as “war” (perhaps referring to the War on Drugs), “dispatch” and “general”. After the homicide, attention shifts to the most visible signs of cartel violence: “body,” “murder,”, “cartel”, “impunity”.

We further test effects on tone of coverage by studying changes in polarity (positive or negative sentiment) as a result of an act of aggression. For the full sample of tweets polarity became more negative afterwards. We observe both a decrease in very positive terms and the appearance of extremely negative language. This is particularly evident in the sub-sample of violence tweets where extremely negative language became prevalent after a homicide. Section A.8 presents the methodology and results in detail.

In one final empirical exercise, we estimate the effect of killings on the grammatical correctness of the accounts' tweets. We do so by estimating the ADTT described in section 4.2, with the fraction of words which are orthographically correct (ie. correctly spelt) as the outcome. The results are shown in Figure A9, and reveal more grammatical mistakes in the months following the killings.

Overall, we find that there are subtle but important changes to the text of the tweets of victimized outlets.

4.5 Long-run patterns

The previous sections showed that targeted outlets reduced coverage in the short to medium run. 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.

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, which include accountants, researchers, psychologists, artists and performers. The sample of analysis is restricted to these occupational categories.

Difference-in-differences: The share of journalists

We implement a long-run difference-in-differences specification to examine whether more violence against the press is associated with changes in the share of journalists. We rely on regressions of the following form:

$$y_{ist} = \alpha + \beta \times violence_s \times post_t + \gamma_t + \gamma_s + \varepsilon_{ist} \tag{1}$$

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 3. 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 almost 15 percent ($\beta = -0.0051$, 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:

$$
\begin{aligned}
y_{isto} = \alpha &+ \beta_0 \times violence_s \times post_t + \beta_1 \times violence_s \times journalist_o \\
&+ \beta_2 \times journalist_o \times post_t + \beta_3 \times violence_s \times journalist_o \times post_t \\
&+ \gamma_o + \gamma_t + \gamma_s + \varepsilon_{ist}
\end{aligned}
\tag{2}
$$

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 4. 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 income. 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.

On the other hand, the decrease in wages reveals that journalists do not appear to be compensated for the increased risk they face. Lower wages could also be explained by better qualified individuals exiting the profession (or moving), which is also consistent with our finding of increased spelling mistakes after killings (Figure A9). That crime affects labor markets more generally in Mexico (including reductions of wages) has also been previously documented in Velásquez (2020).

5 Conceptual framework

We develop a simple formal model of media workers in a context of violence which helps us frame our results and extend our analysis. The qualitative evidence suggested that reporters vary in how they repond to threats of violence (Relly and González de Bustamante, 2014). In particular, some journalists would feel dissuaded from reporting due to fear, while others would see violence or threats of violence as a signal of how important their work is, and therefore feel more commited to it. At the same time, some journalists may adapt to the risk environment by changing the way the news is covered, while still choosing to report (Dorff, Henry and Ley, 2022).

We assume then that there is a continuum of journalists uniformly distributed over $\theta \in [0,1]$, where θ_i captures journalist i's *resolve*, or tolerance for risk. Journalists choose whether to report or not, $r_i \in \{0,1\}$. In addition, journalists can face generalized violence or risk from reporting $v \in \mathbb{R}$, but can also face targeted violence or threats $v_i \in \{0,1\}$.

The payoff to journalists U is a function of their resolve and of the level of violence they face, and of their reporting decision, such that:

$$U_i = r_i(\theta_i - v - v_i(\theta^* - \theta_i))$$

Generalized violence v reduces the journalists utility from reporting. Targeted violence also dissuades journalists from reporting, unless their resolve type θ_i is above a threshold level θ^* which induces backlash (higher utility from reporting). Journalists choose r_i, whether to report or not, to maximize their utility: in this case, if the expected utility from reporting is negative, then journalists would choose not to report.[18]

Our analysis examines the share of media workers actively reporting, defined as M, and the extent to which generalized and targeted violence deters this reporting. Consider first the case in which there is no violence such that $v = 0$ and $v_i = 0$. In this case, all journalists choose to report such that $M = 1$. We now present a set of results based on our model and how these relate to the patterns in our data.

R1: An increase in generalized violence reduces the share of journalists reporting. More pre-

[18] In this simplified framework the decision to leave the profession is collapsed alongside the reporting decision.

cisely $M = 1 - v$, so $\frac{\partial M}{\partial v} < 0$. We refer to this as the *selection effect*.

This result is illustrated in Figure 5, panel (a). As generalized violence increases, individuals leave the profession (selection effect), such that only those with type θ_i higher than v remain. This idea is consistent with our evidence on violence and journalism from census data. We showed that in the most violent states, there are less journalists, and remaining journalists are on average younger and less likely to be married or have kids. This pattern, in turn, supports the idea that some journalists have a higher tolerance for risk: only those with the highest resolve ($\theta_i > v$) continue to work in a setting of increased violence.

R2: If the level of generalized violence is low, targeted violence has a chilling effect. In particular, the share of journalists reporting when faced with targeted violence is $M^T = 1 - \frac{v+\theta^*}{2}$. Consider the case in which we move from an environment in which no journalists are targeted (C) to a threatening environment in which journalists are targeted (T). We define the chilling effect as $\Delta M = M^C - M^T = \frac{\theta^* - v}{2}$, the mass of journalists who exit, or choose to no longer report, as a result of targeted violence (a positive chilling effect indicating that there is less reporting). This result is illustrated in Figure 5, panel (b).

Using an event-study framework following recent advances in the difference-in-differences econometrics literature, our study showed that, over the past decade, journalists killings had on average a chilling effect. In particular, we observe reduced activity from the victimized media outlets in the months after the killings took place.

R3: As generalized violence increases, the chilling effect of targeted violence decreases. That is, $\frac{\partial \Delta M}{\partial v} < 0$. As generalized violence increases, the selection effect causes only the most committed journalists (high θ) to remain in the profession. This, in turn, makes targeted violence less effective as a censorship tool (see Figure 5, panel (c)): when generalized violence is higher ($v' > v$), the chilling effect becomes smaller.

This theoretical result is also consistent with what we observe in the data. Using a difference-in-differences framework with heterogeneous treatments across time, we showed that the killings became less effective over time in inducing self-censorship. In other words, media outlets became less likely to reduce activity after the killings.

R4: If the level of generalized violence v is above the backlash threshold θ^* then targeted violence increases reporting. In the most violent or risky settings (when $v > \theta^*$), the chilling effect becomes negative ($\Delta M < 0$), that is, targeted violence induces an increase in reporting, or a backlash effect instead (Figure 5, panel (d)).

One potential implication of this theoretical result is that the effects may be heterogeneous across states in Mexico, depending on their level of generalized violence (a proxy for v). We therefore bring this additional testable hypothesis to our data. Using data on homicide rates across Mexican states for the beginning of our sample (year 2010), we explore this idea in Table 2. We observe some suggestive evidence that, consistent with the model, the chilling effect of targeted violence is greater in the least violent states. In addition, the estimated coefficients are positive for killings in the most violent states (though not statistically significant), which can be interpreted as evidence of backlash. The estimates are also consistent with an increase in v over time in the initially least violent states, in line with our previous evidence and with R3.

This simple model captures many of the features both from the context of journalism in Mexico (Relly and González de Bustamante, 2014) as well as from what we observe in our data, and it helps to further give our work a broader perspective. In particular, it suggests that the effects of violence against journalists on media activity will depend on the environment these journalists face, which may be changing over time depending on the context.

6 Discussion

This paper studies how the news media in Mexico was affected by the killings of journalists and other media workers. We documented that a majority of victimized journalists had covered crime related news before they were attacked. While 97 percent of these murders remain unsolved, press reports suggest that drug-trafficking organizations planned and carried out a majority of 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.

We observe large reductions in media activity following the killing of an affiliated journalist. Our analysis reveals sustained reductions, of around 25 percent, in victimized outlets' Twitter activity following the murder of an affiliated worker. Different mechanisms might explain the response to these attacks (Pan and Siegel, 2019). The volume of coverage might decrease after a killing because fewer journalists are reporting for a given outlet, which may indicate a *mechanical* effect. At the same time, outlets may change their publishing behavior (*behavior effect*) in response to the attack. More specifically, one objective of the criminals is likely to deter future publishing of certain events through fear. It should be noted that our measured effects combine both mechanical and behavioral responses to the killings.

Importantly, we find that the killings have become less effective in reducing coverage over time, and that outlets in violent states were less responsive to the murders. Analyzing the text of the published tweets, we also find that 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. In the online appendix we also show that journalists in the same network and state as the victim (thus likely more at risk than other journalists) reduced their Twitter activity, suggesting that there may be spillovers to the rest of the media community.

In the long run, these homicides also appear to affect journalism more broadly. 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. These findings reveal that violence has significantly transformed the profession. Furthermore, these patterns suggests that *selection* out of the profession is a key component of these dynamics. That our effects become smaller over time, and that they are initially concentrated in the least violent states, suggests that journalists who remain in this line of work are highly committed, more risk tolerant, and frankly, brave. The conceptual framework we present highlights precisely this previously overlooked aspect of violence against the media.

At the same time, our empirical findings point to violence leading to a degrading in the quantity and quality of the media: reduced activity, younger workers with less pay, and more grammatical errors in posts. Limiting the flow of information, particularly local information that is unlikely to be reported elsewhere, undermines democracy and could further reduce the incentives of public officials to combat organized crime. While the killings have become less effective over time, and despite the commitment of some media workers, the overall pattern that we document is predominantly negative for freedom of the press. Violence against journalists is an effective censorship tool. Given the global prevalence of such attacks (Carey and Gohdes, 2021), protecting press freedom should be an important priority of democratic governments, not just in Mexico, but around the world.

References

Adena, Maja, Ruben Enikolopov, Maria Petrova, and Hans-Joachim Voth. 2021. "Bombs, broadcasts and resistance: Allied intervention and domestic opposition to the Nazi regime during World War II." *Working Paper*.

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, Brian Knight, and Ananya Sen. 2020. "Advertising spending and media bias: Evidence from news coverage of car safety recalls." *Management Science*.

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

Cagé, Julia, Nicolas Hervé, and Béatrice Mazoyer. 2022. "Social Media Influence Mainstream Media: Evidence from Two Billion Tweets."

Cagé, Julia, Nicolas Hervé, and Marie-Luce Viaud. 2020. "The production of information in an online world." *The Review of Economic Studies*, 87(5): 2126–2164.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230.

Carey, Sabine C, and Anita R Gohdes. 2021. "Understanding journalist killings." *The Journal of Politics*, 83(4): 1216–1228.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The effect of minimum wages on low-wage jobs." *The Quarterly Journal of Economics*, 134(3): 1405–1454.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–96.

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.

Dorff, Cassy, Colin Henry, and Sandra Ley. 2022. "Does Violence Against Journalists Deter Detailed Reporting? Evidence From Mexico." *Journal of Conflict Resolution*, 00220027221128307.

Ershov, Daniel, and Juan S Morales. 2021. "Sharing News Left and Right: The Effects of Policies Targeting Misinformation on Social Media."

Foos, Florian, and Daniel Bischof. 2022. "Tabloid Media Campaigns and Public Opinion: Quasi-Experimental Evidence on Euroscepticism in England." *American Political Science Review*, 116(1): 19–37.

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

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*.

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.

Hatte, Sophie, Etienne Madinier, and Ekaterina Zhuravskaya. 2021. "Reading Twitter in the Newsroom: How Social Media Affects Traditional-Media Reporting of Conflicts." *Available at SSRN 3845739*.

Hernández, Anabel. 2012. *Los señores del narco.* Grijalbo, Mexico City.

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.

Krakowski, Krzysztof, Juan S Morales, and Dani Sandu. 2022. "Violence against politicians, negative campaigning, and public opinion: Evidence from Poland." *Comparative Political Studies*, 00104140211066211.

Mazzaro, Kyong. 2021. "Anti-Media Discourse and Violence Against Journalists: Evidence From Chávez's Venezuela." *The International Journal of Press/Politics*.

McMillan, John, and Pablo Zoido. 2004. "How to subvert democracy: Montesinos in Peru." *Journal of Economic perspectives*, 18(4): 69–92.

Morales, Juan S. 2021. "Legislating during war: Conflict and politics in Colombia." *Journal of Public Economics*, 193: 104325.

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

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.

Sant'Anna, Pedro HC, and Jun Zhao. 2020. "Doubly robust difference-in-differences estimators." *Journal of Econometrics*, 219(1): 101–122.

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.

Sun, Liyang, and Sarah Abraham. 2020. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*.

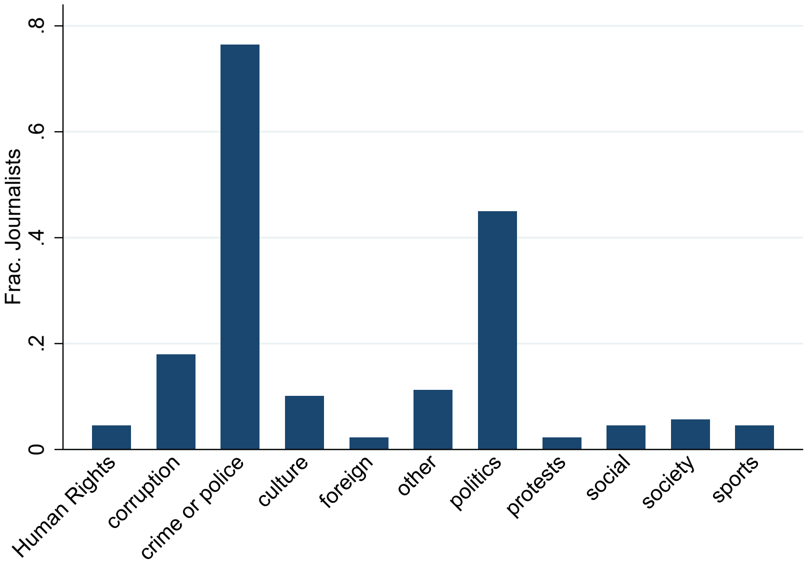
Taddy, Matt. 2013. "Multinomial inverse regression for text analysis." *Journal of the American Statistical Association*, 108(503): 755–770.

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

Velásquez, Andrea. 2020. "The Economic Burden of Crime Evidence from Mexico." *Journal of Human Resources*, 55(4): 1287–1318.

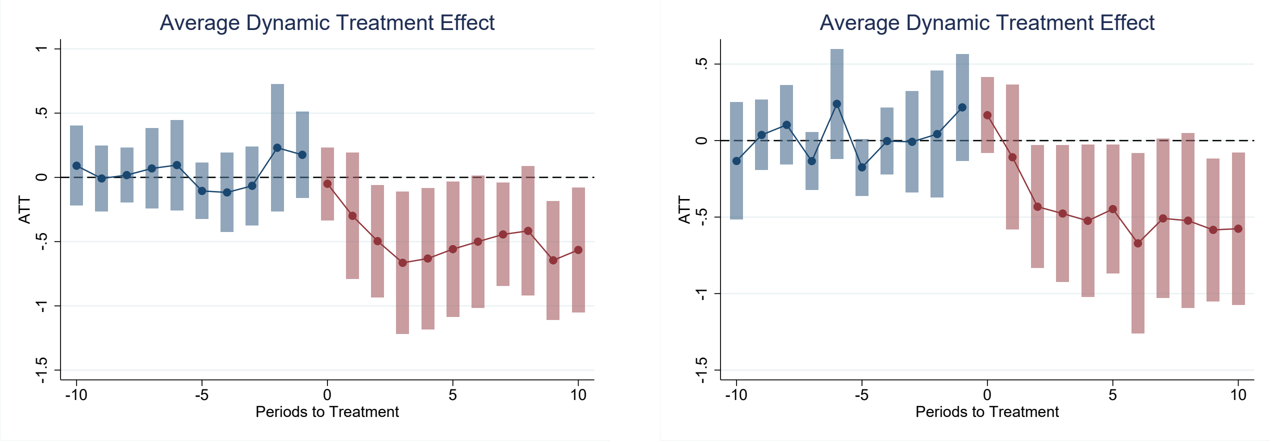
7 Tables and figures

Figure 1: Subjects covered by victims



Note: Own construction based on reports from CPJ as of March, 2022.

Figure 2: Direct effect of murder on volume of tweets

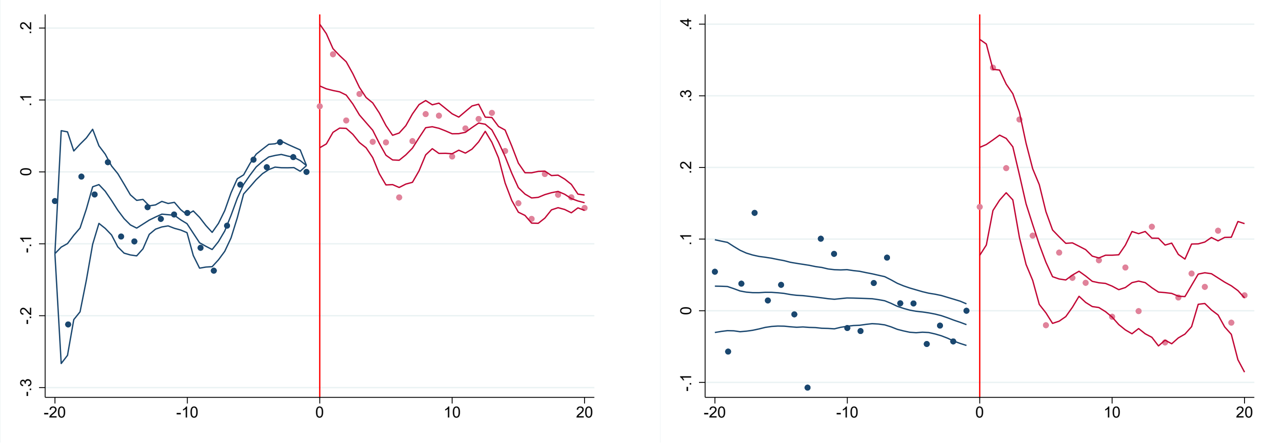


(a) All

(b) Violence tweets

Note: these figures present the point estimates of the average treatment effect of the homicide of a journalist on volume of tweets on those outlets that the victim was affiliated with. The estimate is a weighted average of all difference-in-difference estimates between *treated* (victimized) and not-yet-treated, following Callaway and Sant'Anna (2021) and Sant'Anna and Zhao (2020). Calendar month fixed effects are included.

Figure 3: Google Trends search volume

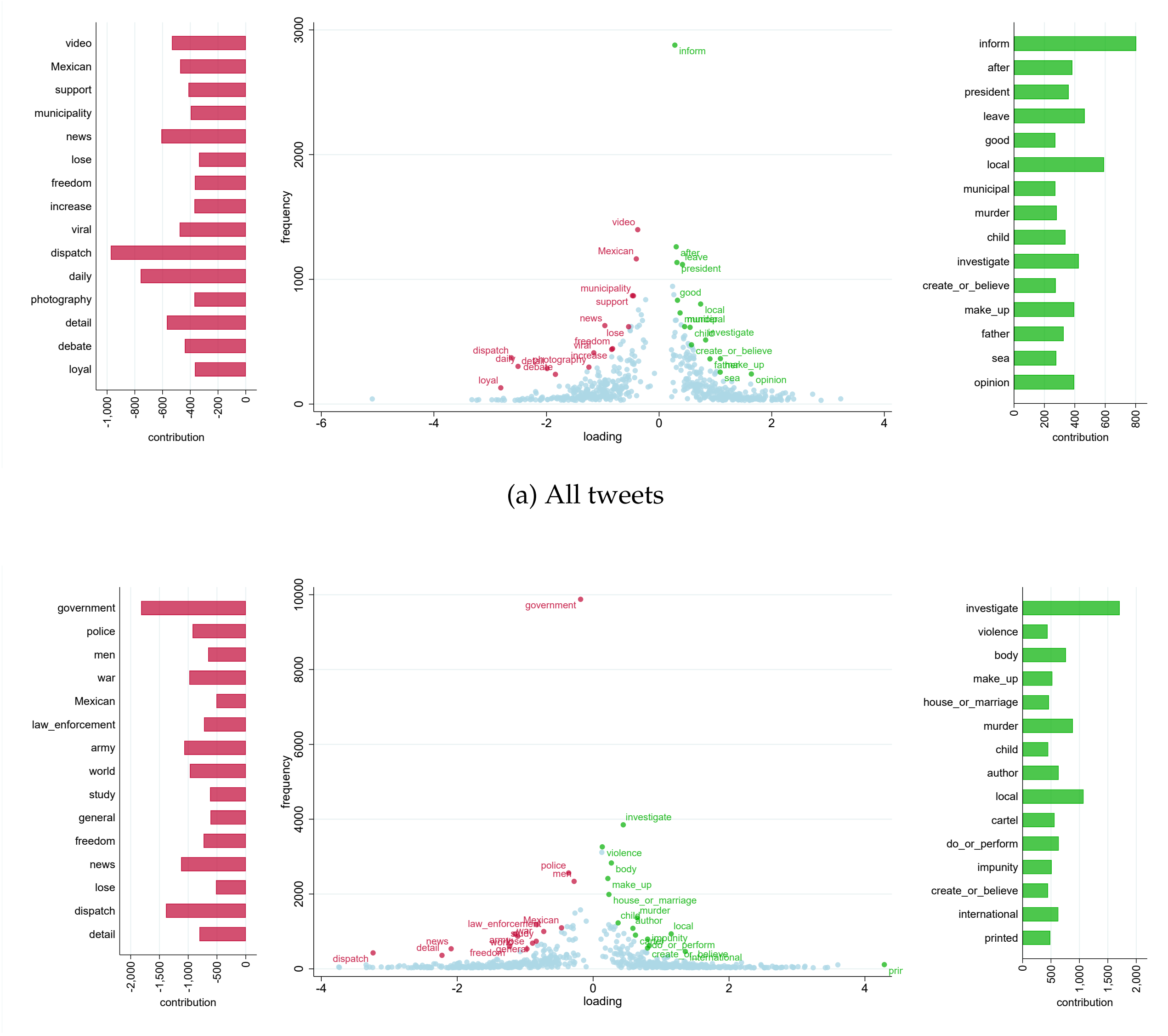


(a) Keyword: *murder*

(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 4: Terms that most predict timing of a tweet

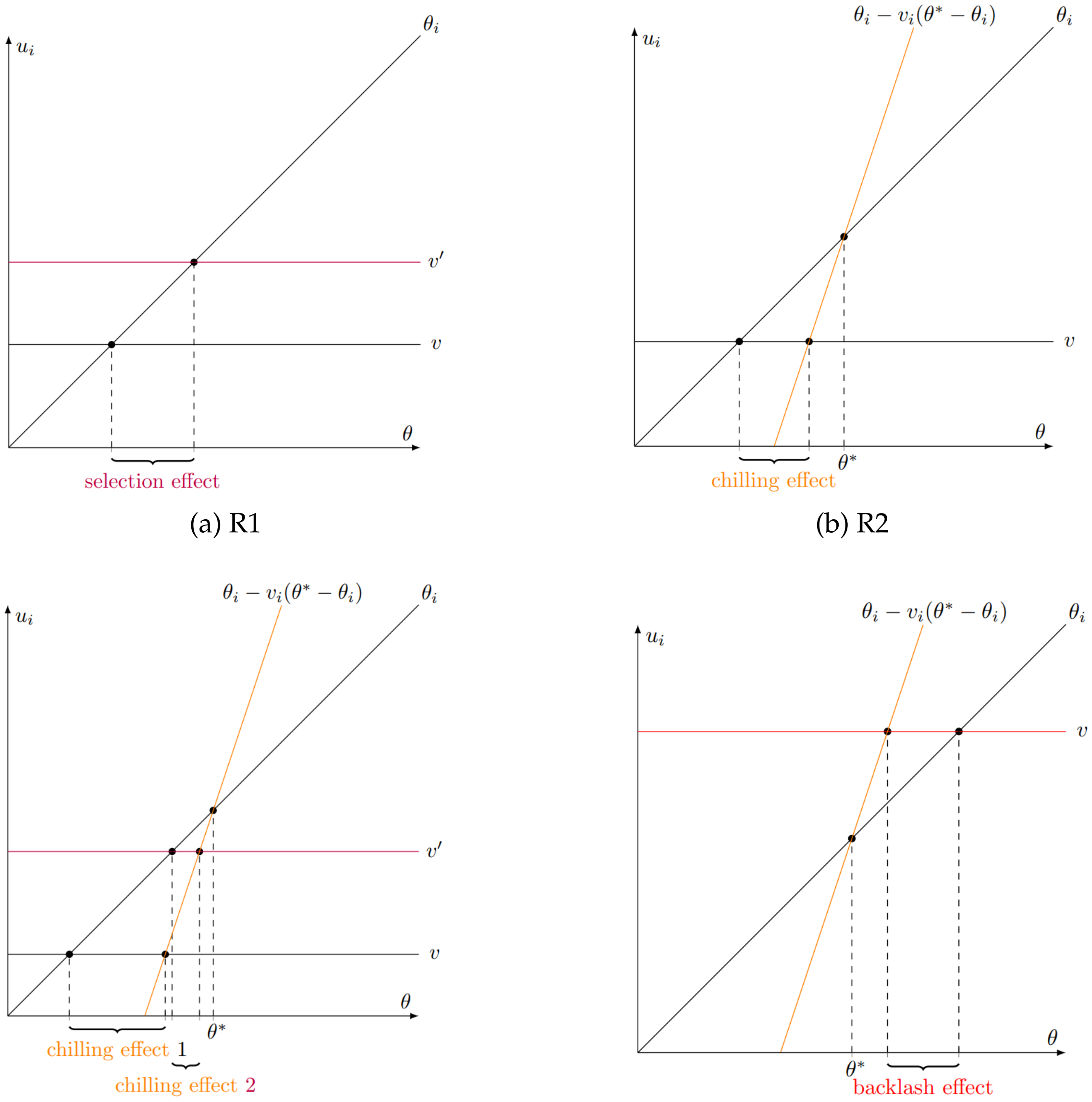


(a) All tweets

(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 non-zero 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 5: Theoretical framework



(a) R1

(b) R2

(c) R3

(d) R4

Note: The figures illustrate the results from the theoretical framework in section 5. The x-axis represents journalists type θ and the y-axis represents the components of the individuals utility function.

Table 1: Effect of journalists killing on Twitter activity, heterogeneity over time

	All tweets					Violent-news				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Post (6-months)	-0.610**	-0.539*	-0.890***	-1.122***	-0.891***	-0.277*	-0.332**	-0.393**	-0.475**	-0.475**
	(0.301)	(0.291)	(0.312)	(0.379)	(0.320)	(0.153)	(0.156)	(0.176)	(0.221)	(0.194)
Post x Yrs. since 2009	0.076**	0.074*	0.110***	0.154***	0.121**	0.034*	0.042*	0.049**	0.075***	0.059*
	(0.037)	(0.039)	(0.039)	(0.046)	(0.047)	(0.019)	(0.022)	(0.023)	(0.027)	(0.030)
N	253255	253254	103018	103018	103018	253255	253254	103018	103018	103018
N-outlets (clusters)	91	91	59	59	59	91	91	59	59	59
Outlet FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Calendar-month FE	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Calendar-month x State FE	No	Yes	No	No	No	No	Yes	No	No	No
Restricted sample	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Event window dummy	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes

Notes: The outcome is the log of number of tweets published for all tweets (columns 1-5) and for tweets about violent news (columns 6-10). Robust standard errors are clustered at the media outlet level in parentheses. Significance levels shown below *p<0.10, ** p<0.05, ***p<0.01.

Table 2: Effect of journalists killing on Twitter activity, heterogeneity by state-level violence

	Most violent states in 2010			Least violent states in 2010			All states	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post (6-months)	0.095	0.257	-0.271	-0.204*	-0.472*	-0.734**	-0.211*	-0.501*
	(0.159)	(0.424)	(0.492)	(0.113)	(0.255)	(0.345)	(0.113)	(0.267)
Post x Yrs. since 2009			0.049			0.080*		
			(0.063)			(0.043)		
Post x Violent-State							0.292	0.736
							(0.190)	(0.484)
N	119677	43929	119677	133578	48847	133578	253255	92776
N-outlets (clusters)	42	38	42	49	42	49	91	80
Outlet FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Calendar-month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time period	All	<2015	All	All	<2015	All	All	<2015

Notes: The outcome is the log of number of tweets published for all tweets. The sample for most violent states are the outlets located in states with above median level of homicides in 2010. Robust standard errors are clustered at the media outlet level in parentheses. Significance levels shown below *p<0.10, ** p<0.05, ***p<0.01.

Table 3: Relationship between violence against journalists and share of journalists

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
post x Nr. MW murd.	-0.0009**	-0.0009**			-0.0011**	-0.0011**		
	(0.0004)	(0.0004)			(0.0005)	(0.0004)		
post x Any. MW murd.			-0.0050	-0.0051			-0.0061	-0.0063
			(0.0046)	(0.0045)			(0.0050)	(0.0050)
N	105281	105281	117673	117673	90639	90639	101421	101421
N-clusters	26	26	32	32	26	26	32	32
State FE	no	yes	no	yes	no	yes	no	yes
Year FE	no	yes	no	yes	no	yes	no	yes
Only wage earners	no	no	no	no	yes	yes	yes	yes

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

Table 4: 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.	-0.0341	-0.0113*	-0.0111**	-0.0685	-0.0033***	0.0037	-0.0197**
	(0.0334)	(0.0060)	(0.0049)	(0.0801)	(0.0010)	(0.0049)	(0.0095)
post x log hom. x journ.	-0.1200	-0.0272	-0.0242	0.2201	-0.0133*	-0.0274	0.1343**
	(0.1483)	(0.0398)	(0.0391)	(0.7352)	(0.0074)	(0.0334)	(0.0497)
N	105281	105281	105281	105281	105281	105281	90639
N-clusters	26	26	26	26	26	26	26
State FE	yes	yes	yes	yes	yes	yes	yes
Year FE	yes	yes	yes	yes	yes	yes	yes

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

Appendix: For Online Publication

A.1 Overall violence and attacks on the press

We test whether there was an unusually high number of homicides among the general population[19] 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 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:

$$homicides_{set} = \gamma_{se} + \gamma_t + \sum_{k=-6}^{6} \beta_k \times monthsSinceKilling_{set} + \epsilon_{set} \tag{3}$$

$homicides_{set}$ 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 windows with respect to the attack.[20] 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}.

Figure A4 presents the β-coefficient estimates from model 3 considering the state where

[19] 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.

[20] 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 six-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.

the media worker homicide *occurred* and the municipality where targeted outlets are *located*. In both specification we fail to find significant changes in general population homicides following an attack on the press.

A.2 Indirect effects

In this section we examine attacks' indirect effects on non-victimized 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.[21] 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 (Halberstam 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 A13 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:

$$\begin{aligned} y_{jsft} &= \sum_{k=-6}^{12} \beta_k \times periodsSinceKilling_k \times inState_s \times inFriends_f \\ &+ \lambda_j + \lambda_{sm} + \lambda_{fm} + \epsilon_{jsft} \end{aligned} \tag{4}$$

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 × month (λ_{sm}) and fixed effects for victimized-outlet follower × 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

[21] Figure A14 shows the Twitter network for the accounts in our dataset.

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*). These tentative results suggest that the killings are effective in reducing coverage not only by the victimized outlets but also by their nearby peers.

A.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. Engagement is defined as the log of likes and retweets (as in Morales, 2021) normalized by the number of tweets, for an outlet-and-day combination. Figure A6 shows a 80 percent decrease in engagement for (or 25% when we control for month and state fixed effects). None of the points estimates are however significant. Results are reversed when we consider the rely on rolling control groups: we measured a not significant 6 to 8% increase in engagement (see columns 3 and 4 of table A3).

The evidence thus is mixed regarding the impact of violence on engagement. It's worth highlighting that change in content may also have an independent impact on engagement.

In section 4.4 we show that average polarity 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.

Lastly, we investigate whether the public tries to learn about these incidents by studying whether there are spikes in related Google searches. Figure 3 depicts coefficient estimates for volume of 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 finding 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.

A.4 Online newspaper articles

To further test whether our tweets are a good proxy of publishing activity we used the Google CSE API to find articles articles for 86 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. Due to limitations in data access, we only have CSE data up until October 2017. Out of the 10,740 queries for outlet × thirty-day window × 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. For the most part, the data is highly correlated with the data from Twitter at the outlet month level A2.

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 figure A3, panel a), that EFIC strongly correlates with national homicides in the country, which allows us to test the quality of the CSE data. The time series of articles for the two outlets that are present in both datasets are, however, weakly correlated, which likely indicates issues with the CSE data. To limit these, we consider as a dependent variable in our event-study estimates an indicator for outlet × thirty-day periods in which *any* articles were published.

Figure A8 depicts our estimates. We observe a fall of approximately 0.1 log points in coverage after an attack that peaks three 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. Notably, we observe a regression to pre-attack levels of publishing within 5 months of the attack.

A.5 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.

A.6 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

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

v_i is a random vector of y_i. In practice, we will take $v_i = y_i$. y_i is a $K - \mathit{dimensional}$ 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.

A.7 Text pipeline

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).[22] Lastly, we consider only terms that appear in at least thirty tweets to limit overfitting and prevent some orthographic errors.[23] 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 the sample of violence tweets.

A.8 Tone of coverage

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

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

[23] 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.

coded as an integer between -5 and 5, inclusive.

Figure A11 shows the frequency of terms by polarity, before and after an act of aggression. To highlight changes in polarity we limit the analysis to terms with non-zero loading in the MNIR estimation, since these terms are the ones that have informational content on the timing of a tweet. In panel *a*, we observe a dramatic decrease in terms with polarity of 5, as well as the appearance of terms with polarity of -5. Interestingly, neutral terms (those with polarity zero polarity) also decrease significantly. These basic results hold for the sub-sample of violence tweets (panel *b*). Most notable, however, is the appearance of large number of extremely negative terms (-5 polarity).

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	2012.645	2.497	2012.511	2.5	.134
Male	.6	.49	.607	.489	.718
Age	38.893	13.075	39.522	13.595	.185
Yrs school	14.404	3.102	14.76	3.032	.001
Married	.488	.5	.516	.5	.117
Has children	.358	.48	.358	.479	.987
Christian	.407	.492	.417	.493	.601
Moved state (pr. 5 yrs)	.099	.299	.08	.272	.07
Urban	.905	.293	.917	.277	.283
Wage worker	.737	.441	.734	.442	.839
No income reported	.136	.343	.136	.343	.961
Log income	8.884	.772	8.954	.844	.022
N	1068	.	2816	.	.

Table A2: Comparison of violence tweets and Google CSE hits

	(1)	(2)	(3)	(4)	(5)
Log Google CSE hits	0.3898***	0.3114***	0.3114***	0.1752**	0.1330
	(0.1075)	(0.0549)	(0.0550)	(0.0767)	(0.0824)
N	3001	3001	3001	3001	2710
R-squared	0.096	0.757	0.757	0.772	0.811
Outlet FE	No	Yes	Yes	Yes	Yes
State FE	No	No	Yes	Yes	Yes
Month FE	No	No	No	Yes	No
Month × state	No	No	No	No	Yes

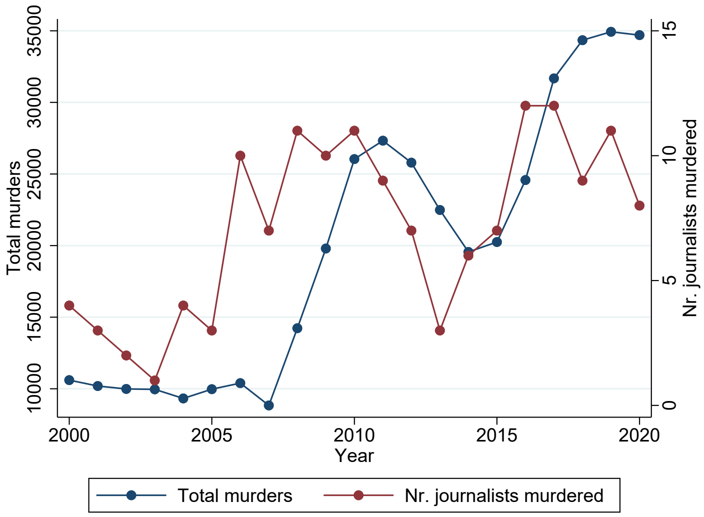
Notes: The outcome is the log of number of violence tweets published for targeted outlets between 2010 and October, 2017. Dependent variable is the log number of articles related to drug-trafficking found by Google Custom Search Engine API on the website of victimized outlets. Robust standard errors are clustered at the media outlet level in parentheses. Significance levels shown below *p<0.10, ** p<0.05, ***p<0.01.

Table A3: Direct effects of an attack (rolling control)

	(1) All tweets	(2) VN tweets	(3) Engagement (All)	(4) Engagement (VN)
Treated=1	0.0077	0.0483	-0.2193**	-0.1179
	(0.1501)	(0.0681)	(0.1010)	(0.1018)
Treated x Post=1	-0.0518	-0.0280	0.0603	0.0820
	(0.1354)	(0.0865)	(0.1454)	(0.1599)
Observations	218632	218632	155232	106348
R^2	0.789	0.759	0.787	0.805

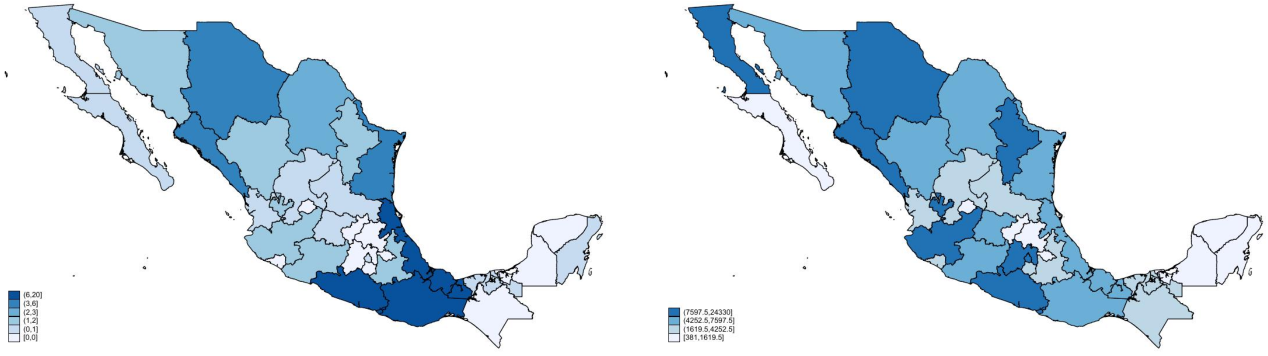
Notes: Robust standard errors are clustered at the media outlet level in parentheses. Significance levels shown below *p<0.10, ** p<0.05, ***p<0.01.

Figure A1: Homicides



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

Figure A2: Homicides in the country (2009-2020)

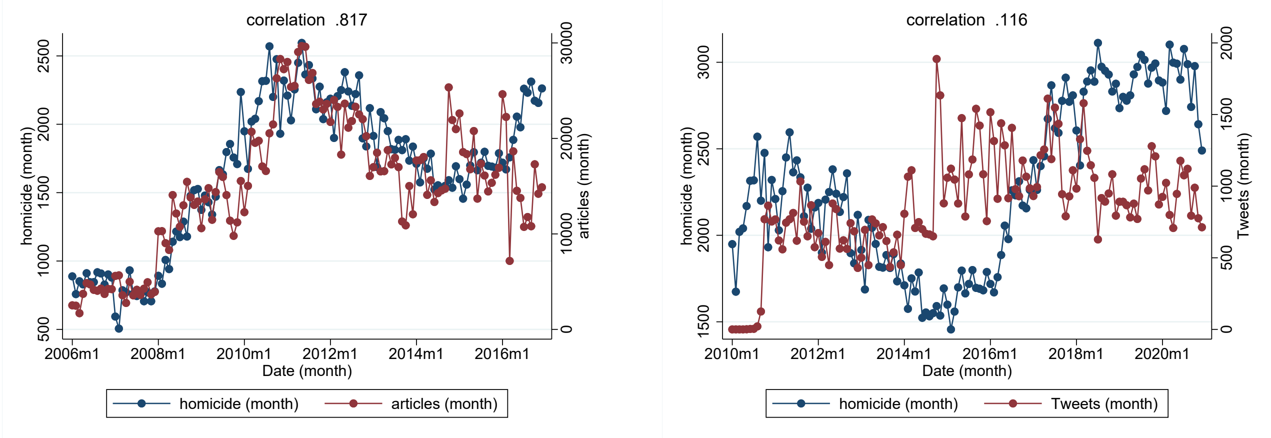


(a) Journalists

(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 2020.

Figure A3: Monthly articles or tweets, and homicides in the country

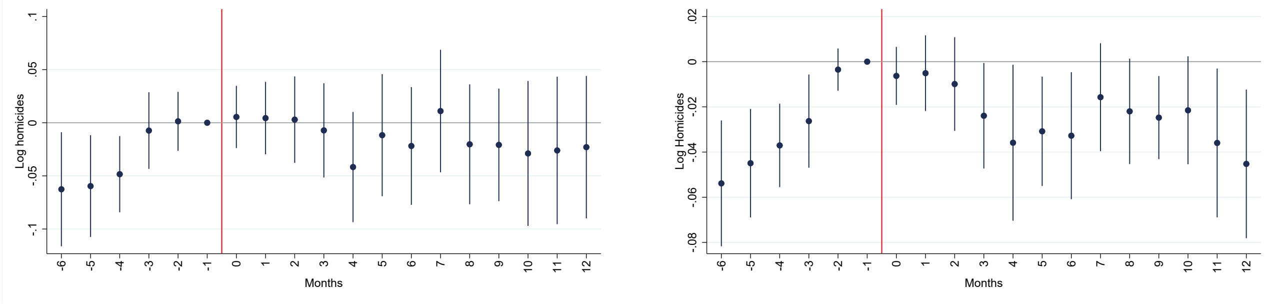


(a) EFIC

(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 3.2).

Figure A4: Homicides around murders of media workers

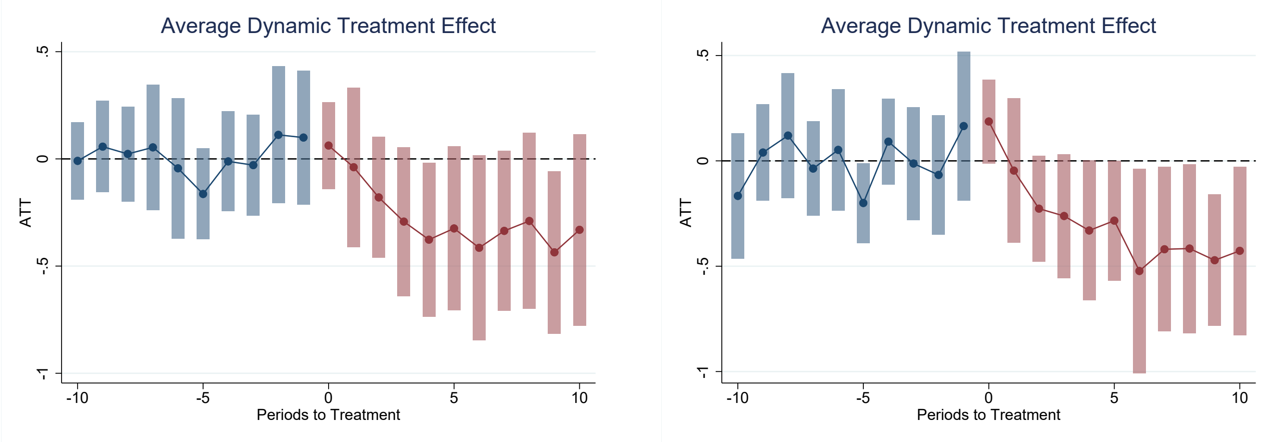


(a) State

(b) Municipality

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

Figure A5: Direct effect of murder (alternative specification)

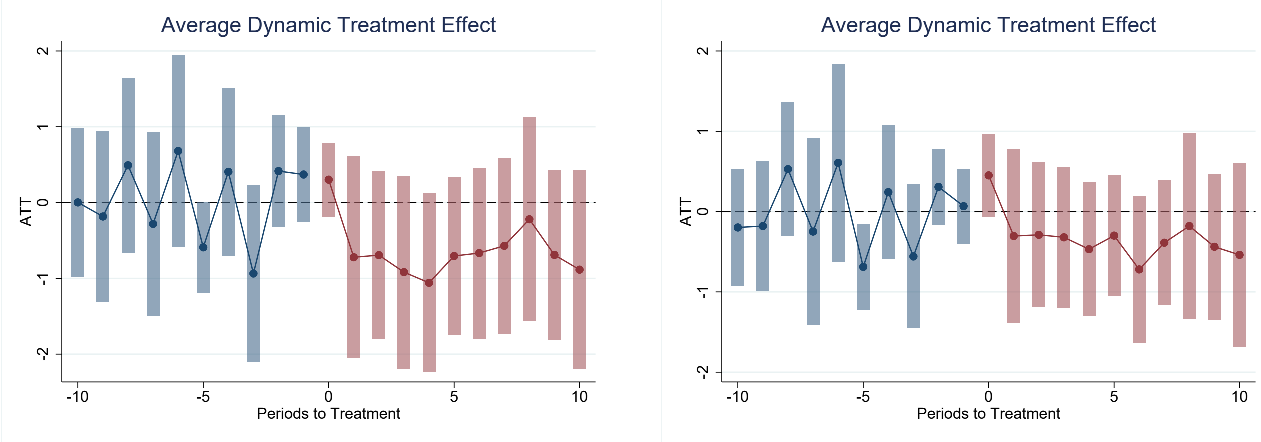


(a) All

(b) Violence

Note: these figures present the point estimates of the average treatment effect of the homicide of a journalist on volume of violence tweets on those outlets that the victim was affiliated with. The estimate is a weighted average of all difference-in-difference estimates between *treated* (victimized) and not-yet-treated, following Callaway and Sant'Anna (2021) and Sant'Anna and Zhao (2020). Calendar month × state fixed effects are included.

Figure A6: Direct effect of murder on twitter engagement

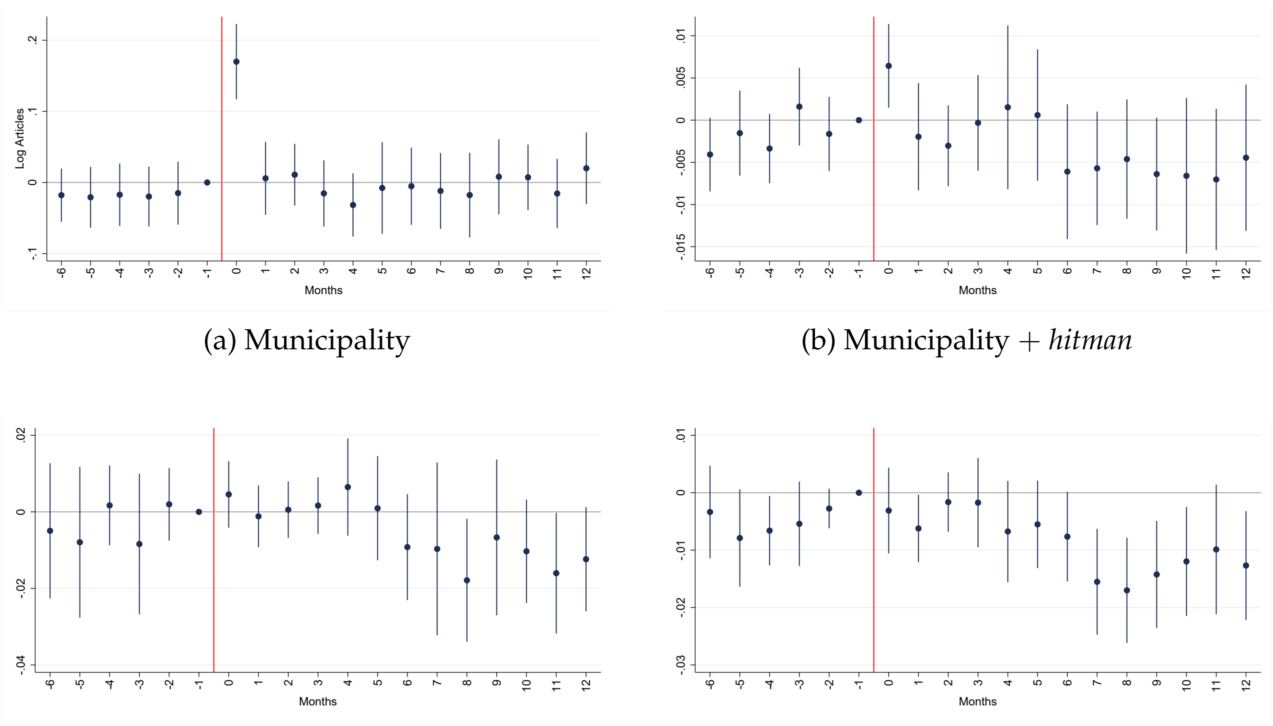


(a) Calendar Month FE

(b) Month × state FE

Note: these figures present the point estimates of the average treatment effect of the homicide of a journalist on average per tweet engagement on those outlets that the victim was affiliated with. The estimate is a weighted average of all difference-in-difference estimates between *treated* (victimized) and not-yet-treated, following Callaway and Sant'Anna (2021) and Sant'Anna and Zhao (2020).

Figure A7: Mentions of municipalities in the national press



(a) Municipality

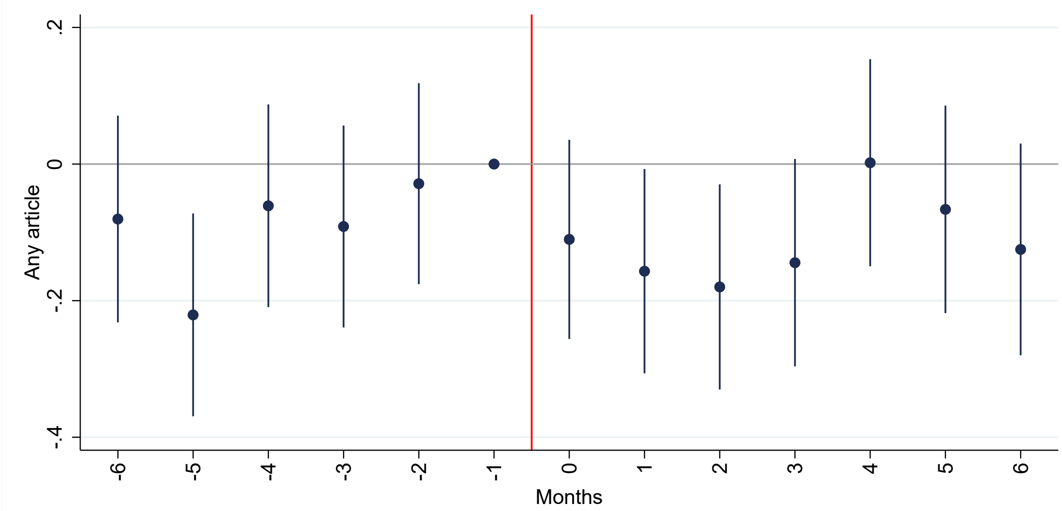
(b) Municipality + *hitman*

(c) Municipality + *Sinaloa Cártel*

(d) Municipality + *Jalisco Cártel*

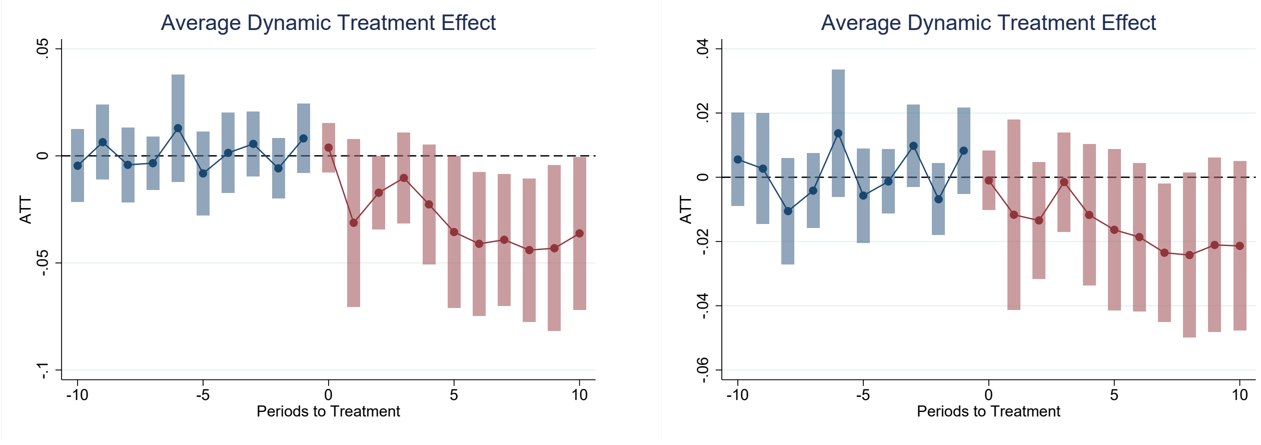
Note: event study considers mentions of municipalities in the national press before and after a journalist was killed there. Panel *b* through *d* consider news items that refer to these municipalities, *along with* an additional term. The Sinaloa Cártel and the Jalisco Organization (CJNG) are the largest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

Figure A8: 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 A9: Fraction of orthographically correct language

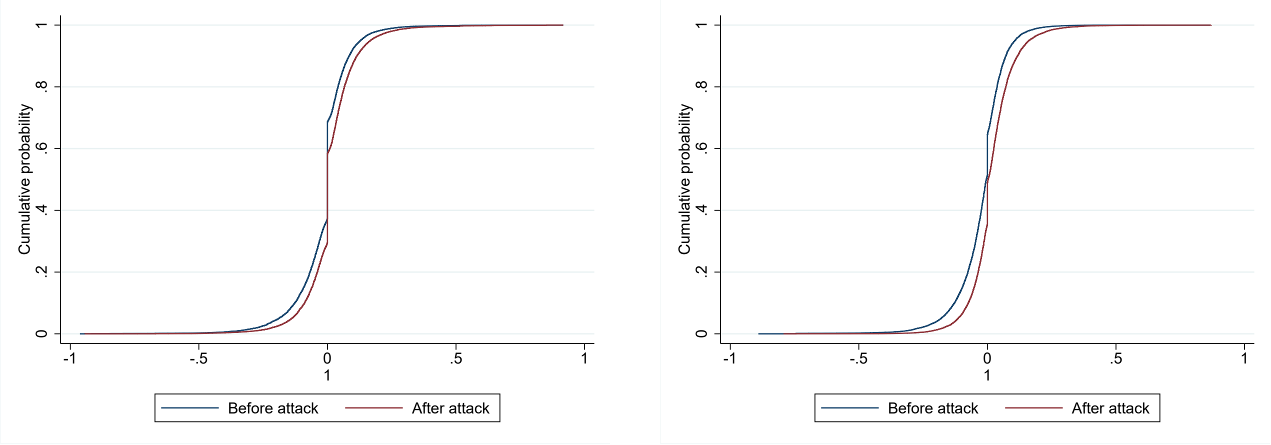


(a) Month FE

(b) Month × state FE

Note: This figures present the Callaway & Sant'Anna estimator for changes in fraction of orthographically correct language in Twitter following the murder of a journalist. The comparison group are not-yet *treated* (i.e., attacked) outlets. Language is classified as orthographically correct if it matches existing terms from the *Real Academia de la Lengua Española* dictionary.

Figure A10: Distribution of sufficient reduction by timing of attack

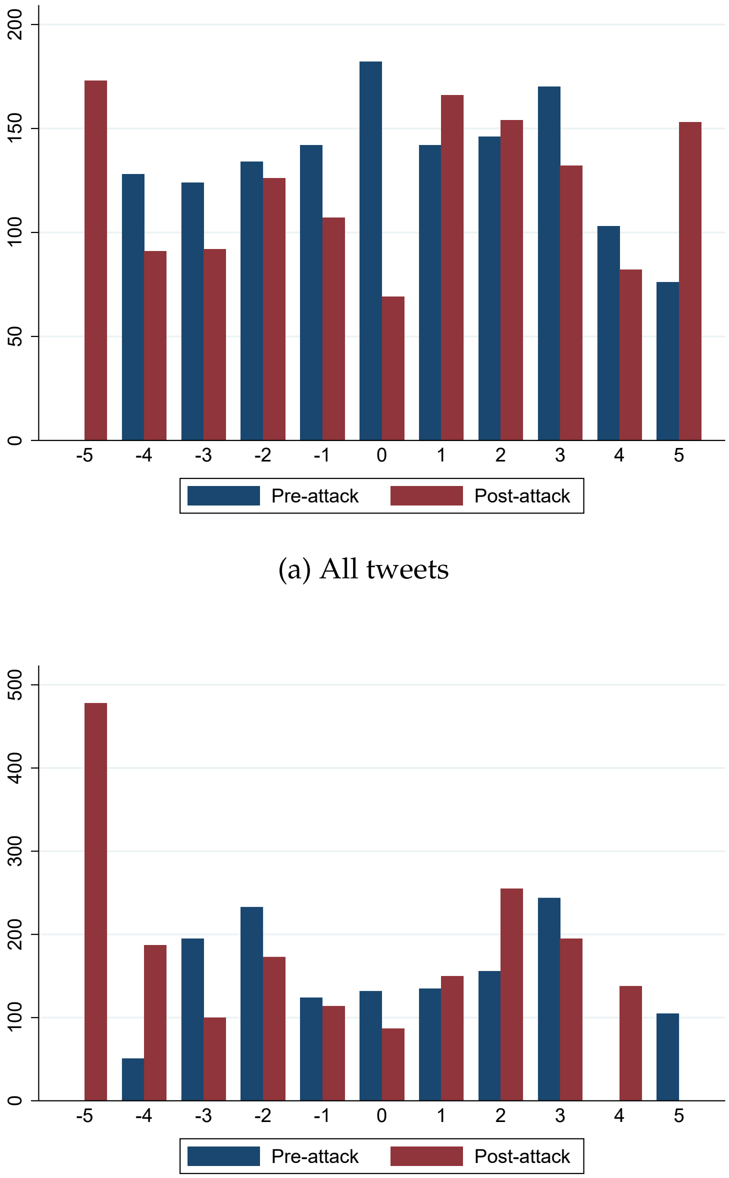


(a) All tweets

(b) Violence tweets

Note: Depicted here is the cumulative 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 A11: Relative polarity before and after an attack

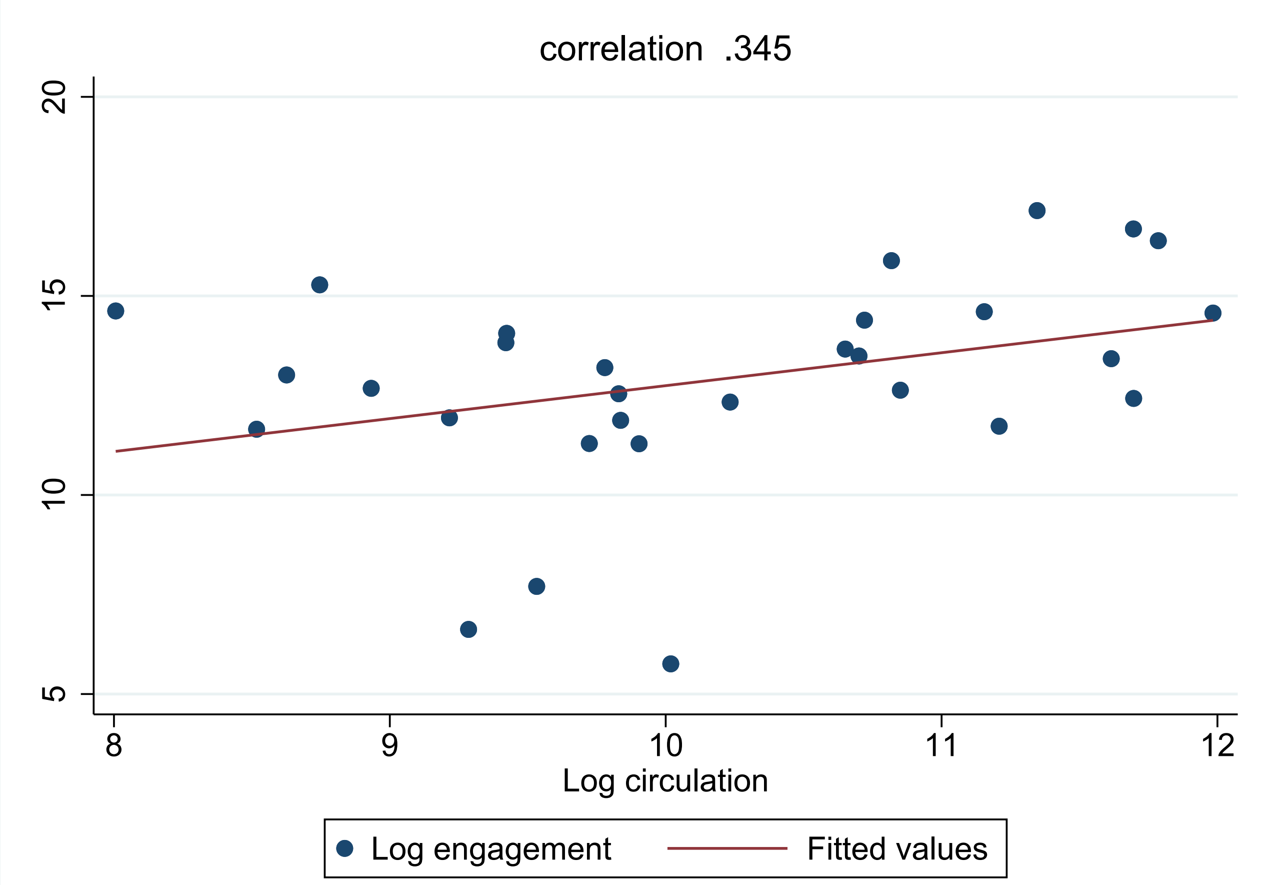


(a) All tweets

(b) Violence tweets

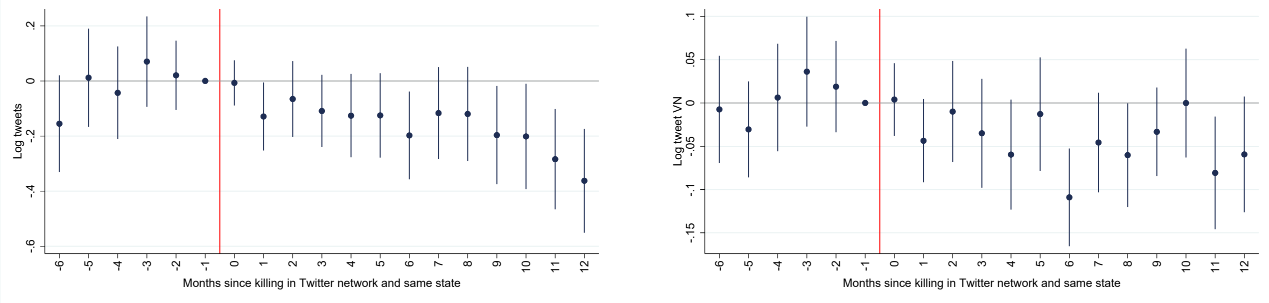
Note: We consider terms with non-zero loading in the MNIR algorithm. These terms are then matched to **?** polarity dictionary and we compute the absolute value of the contributions (defined as loading × frequency) by polarity category (-5, -4, . . . , 4, 5). We considered tweets from victimized outlets 180 days before and after an attack.

Figure A12: 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 A13: Indirect effects on volume of coverage for journalists following victimized outlets on Twitter and located in the same state

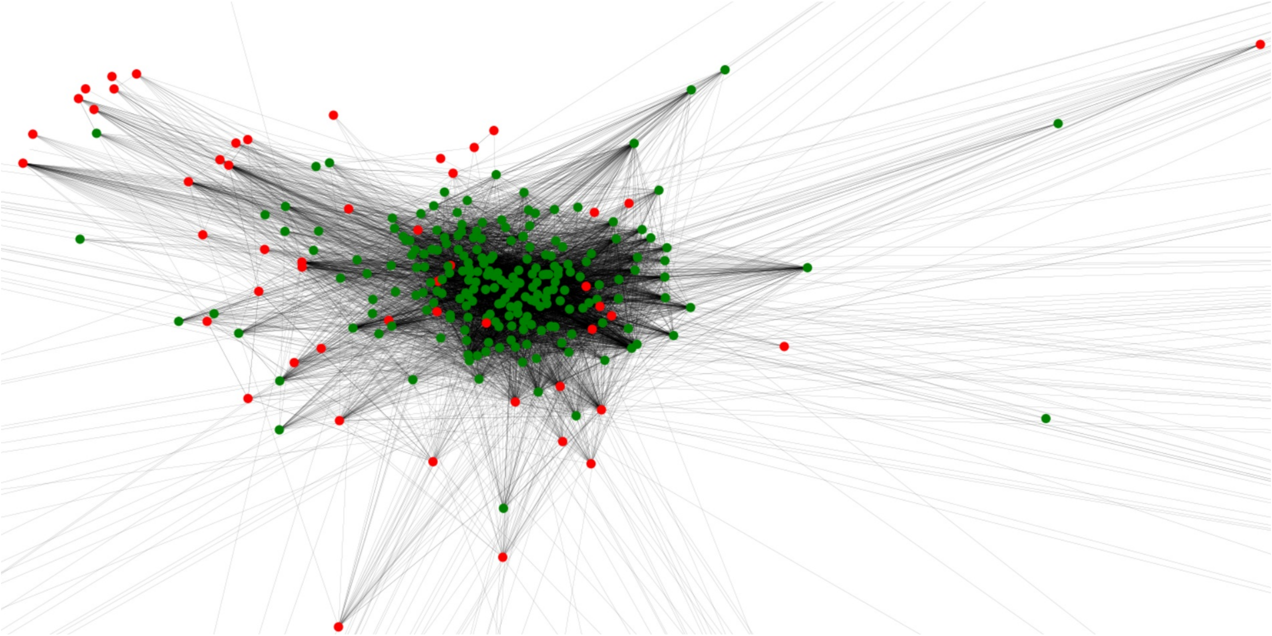


(a) All tweets

(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 A14: 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.