Quantifying and Contextualizing Violent Collective Action Event Datasets

William O'Brochta[*] and Sunita Parikh[+]
Published in *Social Indicators Research*

How does newspaper-based event data compare to a government data source? While scholars have long recognized the importance of and biases present in newspaper-based event data, few studies have compared newspaper reports with official government data to better understand the severity and impact of such biases. We develop this comparison in the context of riots, a form of violent collective action that represents an important middle ground between peaceful protests and protracted civil conflict. Using newly collected police precinct-level government data from India, we compare these data to a high-quality newspaper source. Though similar at the aggregate level, newspaper riot reports correlate poorly with government data at the local level. We model the frequency of newspaper and government riot reports based on literacy, location, and other demographic characteristics to better understand the discrepancies between these two sources. We conclude that newspaper riot data does partially reflect aggregate riot trends, but the newspaper editorial process also plays an important role. Government data is better for within country comparisons and for analyzing event trends over time. Our findings suggest that using collective action event data from both sources may help ensure that results are not driven by biases in either data source.

Keywords: civil conflict, communal violence, newspaper data, India, riots.
Conflict of interest statement: The authors declare no conflicts of interest.
Funding: Weidenbaum Center on the Economy and Public Policy.
Acknowledgements: We thank the audience at the 2018 Midwest Political Science Association Annual Conference, Lee Ann Banaszak, and the editor and reviewer for helpful comments.

[*] Corresponding author: Assistant Professor of Political Science, Department of Political Science, Sociology, and Geography, Texas Lutheran University, 1000 W. Court Street, Seguin, Texas 78155. wobrochta@tlu.edu.
[+] Associate Professor of Political Science, Washington University in St. Louis.

1. Introduction

Newspapers have long provided scholars across a range of disciplines with data on collective action events. The regularity with which newspapers are published, their attention to political and social events, and their granular level of daily reporting have made articles and feature stories the raw material for the collection, coding, and analysis of events of interest. Even in the age of social media's firehose of information and opinion, few sources are as readily available and as regularly produced. With new machine learning methods scraping and aggregating enormous amounts of newspaper data into new datasets, it is critical to understand how newspaper data is constructed.

For as long as newspapers have formed the basis for social science data, there have been discussions of the potential drawbacks of using them as a relatively unbiased source. Scholars have generally taken the position that as long as we are aware of the risks, newspapers are superior in quality and availability to other data sources. This position has been bolstered by the difficulty of acquiring equivalent data from official sources, especially when conducting national or cross-national research in settings where data are not centrally collected or disseminated or where official repositories fail to sort and categorize data according to scholars' categories of interest. But studies comparing newspaper and government data have repeatedly uncovered discrepancies in the number and type of events reported in each source. Newspaper-based datasets tend to report events which are larger, feature police presence, and occur in concert with narratives that are concurrently in the news (Andrews and Caren 2010; Barranco and Wisler 1999; Oliver and Many 2000; Oliver and Myers 1999; Snyder and Kelly 1977).

Despite the prevailing argument that the advantages of using newspaper data outweigh the drawbacks, empirical evidence challenging this consensus has been available for some time.

In a 2004 study, Myers and Caniglia (2004) find that newspaper coverage is affected by event size, social issues in vogue at the time, media reporting priorities, seasonal patterns, violence intensity, and political orientation. This research corroborates the findings in Earl et al. (2004)'s review of the literature on newspaper reports of protest events, where they distinguish between selection bias and description bias. They note that while "'hard news' is mostly subject to errors of omission, 'soft news' (i.e., impressions and inferences of journalists and commentators) is subject to multiple sources of bias" (Earl et al. 2004, 72). As soft news becomes more prevalent in print journalism and thematic reports supplant detailed descriptive reporting, the impacts of newspaper bias are increasing. "Newspapers are not a transparent conduit of information about protest, and that important systematic selection factors and processes affect the types of data available in newspapers" (Ortiz et al. 2005, 398). Echoing Earl et al.'s review, Ortiz et al. (2005) address selection bias and argue that its effects must be taken more seriously. Initial decisions by media that introduce selection and description bias are compounded by scholars who contribute another dimension of bias through their sampling techniques and source selection decisions.

Yet, while sociology has been at the forefront of social science disciplines which use and critically examine newspaper records as primary data (Abbott 1995; Olzak and Shanahan 2003), little existing work directly compares newspaper data to their most readily available alternative, government data, to analyze how these data differ and the tradeoffs in using them. Doing so is important because scholars need to fully understand the advantages and disadvantages of each dataset to make informed decisions about how to employ these data in their work. A comparison is also broadly methodologically useful since collective action events continue to be prevalent worldwide. Scholars increasingly interested in understanding collective action can benefit by

reflecting on the data generating process that produces all forms of event data, particularly that from newspapers and government collected data.

We examine differences in newspaper and government data on riots, a form of violent collective action particularly likely to attract media attention. We take as our case riots in India reported in the *Times of India* newspaper compared to newly collected official government data on riot prevalence. Our comparison starts with a detailed description of how each riot dataset is generated, as the data generating process impacts what is counted as a riot and the resulting dataset. When then examine the correlation between these data at different levels of aggregation. While we observe similarities in riot reporting at the aggregate level, at the state and district level we observe high numbers of reported riots in the government dataset whereas the newspaper data records few if any newspaper articles about riots. This finding leads to an additional analysis where we seek to explain both newspaper and government riot reports using past riot reports and local district-level characteristics. We find that newspaper reporting exhibits selection bias, while also partially reflecting trends in government reporting. Given these findings, we conclude by suggesting ways in which each data source can be utilized most accurately and effectively.

Our aim is to provide scholars of collective action with guidance regarding the appropriate uses of different kinds of collective action data. In doing so, we move beyond recent work that compares different sources of newspaper data to each other and instead think more broadly about the construction of collective action datasets and these dataset's potential uses. As records are increasingly digitized, understanding and describing how data are generated and their best uses becomes even more critical (Johnson, Schreiner, and Agnone 2016).

2. Case Selection

We take as our case the prevalence of riots in India in the latter half of the 20th century. Since India gained independence in 1947 the government has struggled to quell all types of political violence. Riots occupy a middle ground between non-violent collective action (e.g., non-violent protests and demonstrations) and rare violent events such as lynchings and civil wars (Hagen, Makovi, and Bearman 2013; Weidmann 2015). Riots --- defined here as an unlawful assembly using violence --- include collective violence, but they are not always deadly, and they range from very small one-off events to large-scale collective action which can comprise dozens or even hundreds of smaller events (Indian Penal Code Section 146). The focus of collective action comparisons between newspaper and government data has primarily been at these opposite ends of the spectrum. Our strategy of assessing different approaches to riot data collection and analysis provides a bridge between these categories by analyzing a form of regularly occurring collective action which has the political implications and frequency of protests and the potential deadliness of civil wars.

Rioting killed hundreds of thousands during the partition of British India (Brass 2003). Scholars over the years have sought to explain the causes and consequences of these riots, and the persistence of such events has made India the most common locale to study violent collective action. But although there is agreement over the importance of understanding the role of violence in Indian politics, there is no corresponding consensus on how to measure it. The vast majority of studies have been qualitative, beginning with Lambert (1951) and continuing through articles and books to the present. Bayley (1963) offered a tentative foray into quantifying Indian riots,

noting as he did so the selection biases inherent in newspaper-based riot reports.[1] Nayar (1975) offered a subsequent analysis using National Crime Records Bureau data, but it was disaggregated only to the state level. Between Nayar's 1975 book and 1996, there were no comprehensive quantitative studies of riots, even as qualitative accounts flourished in both scholarly and journal publications.

The newspaper-based dataset on riots collected by Ashutosh Varshney and Steven Wilkinson, analysis of which was first published in 1996, has become one of the most widely used sources for scholars studying religious and ethnic conflict in India (Varshney and Wilkinson 1996). Hundreds of scholarly articles have used these data, both as an independent and dependent variable. The dataset has been downloaded over 125 times in the last three years.[2] Painstakingly hand-collected from newspaper articles on riots as reported in the national edition of the *Times of India*, the dataset collects riot events from 1950 to 1995. Every riot event that was coded as having a Hindu-Muslim aspect *and* that resulted in at least one death was recorded in the database.

While Varshney and Wilkinson took care to determine that their measure was not unusually biased, news coverage even of large riot events is shaped by the perceptions of individual reporters and editors, as well as by their expectations about what coverage is most desired by their audiences. In the Indian case, frames that emphasize Hindu-Muslim conflict can dominate discussions of events even when they are not the precipitating factors or when other

[1] The number itself is difficult to estimate since scholars cite either the original dataset, Varshney (2003), or Wilkinson (2004).
[2] The initial dataset covered the period from 1960-1993. Version 2 of the dataset, which superseded the initial version and is the most widely used, covers 1950-1995. We cite the authors' 1996 discussion of data choice and coding because these data choices and coding rules appear to have guided later collection and analysis.

motivations are present (Sonwalkar 2004). These factors arise not only in coverage of major riots but can also occur in reporting on small events in locations with histories of specific caste or communal tensions.

We explore the robustness of the *Times of India* (TOI) data by contrasting it with a newly collected dataset on government reports of riots. Government riot reporting proceeds quite differently from newspaper reporting and is precipitated by the filing of a notification of an event matching the specific legal description at a local police station. These reports are collected yearly by the National Crime Records Bureau (NCRB), and reports are issued as part of the annual *Crime in India*.

We select these datasets for comparison because of their prevalence and longstanding importance in riot analysis. Yet, it is important to acknowledge that the Internet and social media have enabled scholars to collect vast amounts of data, including riot reports. Some such datasets --- like GDELT --- aggregate newspaper reports of events including riots. Other datasets rely on social media posts, images, or videos to determine the presence of a riot and to estimate its size (e.g., Sobolev, Chen, Joo, and Steinert-Threlkeld 2020; Zhang and Pan 2019). Before we can more fully understand the data generating process and reliability of these new data sources (see Demarest and Langer 2022; Hoffman, Santos, Neumayer, and Mercea 2022; Ward, Beger, Cutler, Dickenson, Dorff, and Radford 2013), we argue that we must first understand their underlying components. The reliability of aggregate newspaper reports partly depends on each constituent newspaper. Similarly, social media data measures human behavior and judgement about riot reporting, just as NCRB data relies on individual human riot reports. Our goal in analyzing the similarities and differences across the TOI and NCRB datasets is to identify their relative strengths and weaknesses and to describe situations in which one or the other is better

suited to the analysis being undertaken. Such work also represents the first step toward more systematically analyzing aggregate newspaper and social media data.

3. Sources of Bias in Newspaper Reporting and Government Data

3.1. *Times of India* Data

Varshney and Wilkinson chose the TOI as their data source because the newspaper is nationally published and read. While they used the flagship Mumbai edition, the TOI historically staffed bureaus throughout India with reporters who had local expertise.[3] The authors assert that the TOI is the only newspaper among the major English-language dailies which "a) unlike *The Pioneer*, *The Statesman*, or *The Hindu*, has truly national coverage of Hindu-Muslim violence; b) unlike some other newspapers, has refused to run the potentially most inflammatory stories about communal violence without having them double-checked; and c) is readily available and covers the whole period in which we are interested" (Varshney and Wilkinson 1996, 9).[4] They consider the TOI to be the most respected English language newspaper in India with a good reputation for reliability and coverage throughout India (Wilkinson 2008).

To test whether the Mumbai edition over-reported local riots, Varshney and Wilkinson calculated ratios of riots with one to three deaths over riots with four or more deaths for all states which had both types of riots.[5] They find that the distance from Mumbai is uncorrelated with whether small or large riots are reported, implying "no systematic reporting bias in favor of

[3] This was true during the time period under investigation, from 1950 to 2005.
[4] Here the authors assume that the number of deaths is correlated with riot size.
[5] This essentially means that different states over and under-reported small riots, e.g., West Bengal had a ratio of 1.5 while Bihar had a ratio of 0.7.

Western India was noticeable" (Varshney and Wilkinson 1996, 49).[6] Wilkinson (2008, 279) later equivocates, stating that the TOI probably over-reports small riots near Mumbai.

The TOI's reporting on riots where no deaths occurred "showed a strong bias towards covering riots in Maharashtra and Gujarat and the under-reporting of events of similar magnitude in other states," and the authors therefore chose to record information about injuries and deaths in their dataset (Varshney and Wilkinson 1996, 10). As a result, Wilkinson (2004, 248) recommends using the TOI data in cases where at least one death occurs.

The authors also developed detailed coding rules for determining whether a riot should be classified as a Hindu-Muslim riot. Wilkinson (2004, 255) defines a "communal" riot as one where "there is violence and two or more communally identified groups confront each other at some point during the violence." Since "communal" is also used to refer to other types of inter-religious conflict like Hindu-Sikh or Christian-Muslim, they coded as Hindu-Muslim riots only those events where "the labelling of the riot in the newspaper was supported by the description of the symbols and issues involved" (Varshney and Wilkinson 1996, 10). If an event description included other types of conflict, such as caste or tribal, as well as Hindu-Muslim conflict, it was coded as "probable" in terms of fitting the category (Varshney and Wilkinson 1996, 57-58). In addition, violence between police and a communally-identified group or by two communally-identified groups was counted as "probable" if it preceded or followed Hindu-Muslim conflict (Varshney and Wilkinson 1996, 53-54).

[6] Lalita Kumari v. Govt. of UP, (2014) 2 SSC 1, sections 24-28.

3.1.1. Potential Issues

The 2006 version of the Varshney and Wilkinson dataset is the standard for studying patterns of violent collective action. Nevertheless, scholars have identified several potential problems with the methods used to collect and categorize the data. The first problem is that it is exceedingly difficult to distinguish a "true" or solely communal riot from a riot that at some point takes on a communal nature (Bhavnani and Lacina 2015, 771). Varshney and Wilkinson note this in their coding appendix, offering the 1985 Ahmadabad riots as a case where ongoing violence changed over time from caste-based to communal conflict. The coding rules state that riots of this type are to be coded as "probable" rather than "definite," but scholars using the data have ignored this distinction. This coding decision makes it difficult to say that the TOI dataset captures only Hindu-Muslim riots; instead, it is more likely that it captures all riots in which TOI reporters believe Hindu-Muslim conflict plays a role, including events with other ethnic, economic, or political motivations. This over-inclusion is not necessarily a problem for Varshney and Wilkinson's original research, since their task was to evaluate events where Hindu-Muslim conflict played any role. But it is a problem for studies which use the dataset as a proxy for religious violence more generally or for civil conflict, as the dataset is really an incomplete measure of all Indian riots.

The second problem is that while Varshney and Wilkinson address some of the potential biases of using newspaper reports as their data source like distance from the newspaper office to the event, writers and editorial staff possess un-measurable biases that make quantifying riot reports from newspaper data difficult. This may help explain why Myers and Caniglia (2004) found large differences between newspaper and government riot reporting in the United States.

Scholars have specifically studied the poor quality and non-systematic coverage of the *Times of India*. Sonwalkar (2004) finds that coverage in the Northeast region of India is severely lacking in English language newspapers. More importantly, he uncovers systematic biases that exist in English language newspaper newsrooms where editors decide which stories to publish based on what will be of most interest to readers, not what is newsworthy. This is related to an advertising centered model of journalism that quickly took root after the 1977 emergency (Jeffrey 1993). Indian newspapers, including the TOI, have shifted from pure interest in reporting news to writing news stories to appeal to advertisers and readers in order to increase profits (Sonwalkar 2002); the TOI is willing to print whatever news is necessary to be profitable (Jain 2017, 167-168). Menon argues that in reporting the 1992 riots in Ayodha, "sections of the so-called national English press were prone to reportage that was provocative, [and] that relied on rumor" (Ganguly, Diamond, and Plattner 2007, 185). As a TOI editor told a scholar, "I am told not to put stories of prisons on the first page. No one wants to read them. Although people do want to read them! *But the management wants only feel-good things.* Readers should not be upset, because then the ads do not go down well" (Rao 2010, 150).

Mody (2015) studies newspaper coverage of the Maoist armed struggle for three regional newspapers, plus the TOI and *Dainik Jagran*, the largest circulation newspaper in the world. The study finds large differences between coverage of Maoist events as well as a large discrepancy between the percentage of articles reporting the cause of the incident in the TOI (6.18%) and regional newspapers (23%). Further, topics mentioned in relation to these incidents varied widely across newspapers with the TOI mentioning paramilitary involvement in only 10% of articles, while other newspapers mentioned their involvement in 43% of articles (Mody 2015, 743).

Similarly, O'Brochta (2019) compares riot reporting in Hindi and English newspapers, finding substantial differences in reporting between them.

One suggestion Wilkinson (2008) provides is that any issues with the TOI dataset can be resolved by including newspapers from all regions and all languages in India. While this initially seems appealing, Supplemental Information section 1 (SI.1) compares riot reports in the TOI to three other English language newspapers in India (see also Marsh 1991; Murthy, Ramakrishna, and Melkote 2010). One might expect that trends over time are similar, but no pattern is discernible. This implies that newspapers are selecting to report riots in some non-systematic manner.

Although the TOI may be the best newspaper source for Indian riots, its publishers and reporters make the same kinds of decisions, based on the same kinds of concerns that lead to selection bias in studies based on the *New York Times*, *Washington Post*, and other respected newspapers. Decisions inherent to the news gathering process, such as the media attention cycle, seasonal patterns in protest activity, and the effect of institutionalized violence on perceptions of newsworthiness play a pivotal role. These potential issues with the TOI data mean that those studying violent collective action events should check the robustness of the TOI data with information from other sources. In the next section, we introduce a newly collected measure of riots using government data.

3.2. National Crime Records Bureau Riot Reports

Varshney and Wilkinson created the TOI dataset to address problems they identified as present in government reported riot data. First, they noted that government statistics are usually only made publicly available after questioning of the Home Ministry by members of parliament and

other special investigations. Thus, there are huge gaps in these data representing major riot events where government data are not released to the public. Second, when data is provided, it is often aggregated, making it difficult to identify the precise locations of riots. Third, government data relies on the definition of a "communal incident," which is defined differently by different states and many communal incidents may make-up a single riot. Finally, government reported data is rarely double-checked, so any police officer charged with reporting riot data can apply his or her own definition of a communal riot without later quality control. A less charitable interpretation of this fourth point is that police officers may demand bribes to record riot reports, may not record riots involving well-known figures, and may pay more attention to riot cases where police were injured or victims were already compensated (Wilkinson 2004, 244).

To be clear, there are two types of government riot data available. The first, what we call "Home Ministry Data," is presented in the Indian parliament upon the request of legislators. These data are constructed by parsing through state and district level reports about riots in order to count the number of riots occurring in each district-year. Communal riots are classified separately from riots with other causes. Home Ministry Data (HMD) suffers from erratic releases, aggregated statistics, and inconsistent definitions of riots and communal incidents. Wilkinson uses HMD to cross-check the TOI dataset in Uttar Pradesh (Wilkinson 2004, 249). He finds that HMD fails to mention 58 riots included in the TOI, while the TOI misses only 7 riots included in HMD. We agree that HMD is a poor representation of Indian riot trends.

We focus on data from National Crime Records Bureau (NCRB) *Crime in India* reports. The NCRB is overseen by the Home Ministry, but it is not involved with the collection of HMD. When a riot is ongoing, any witness is able to go to a police precinct and file a First Information Report (FIR). The FIR is a police record keeping system that indicates that a cognizable crime

has occurred. Cognizable crimes are more serious crimes wherein police are authorized to immediately begin an investigation and to use the FIR as the basis with which to enact an arrest. FIR data is collected at the police precinct level and sent to the district and then to the state, where it is aggregated and delivered to the NCRB. The NCRB is responsible for ensuring the validity of these data by comparing the reported number of riots to previous years' data to look for any discrepancies. The NCRB definition of a riot uses the Indian Penal Code definition of "an event involving a group of five or more individuals who are illegally assembled and who use violence in pursuit of a common goal" (Indian Penal Code Section 146).

3.2.1 Strengths and Weaknesses

The NCRB method has a number of strengths. First, the use of a uniform riot definition that is inclusive of all riots means that these data avoid problems with newspaper or researcher interpretation of whether a riot event was truly Hindu-Muslim or if it had some other motivation. Further, because these data are available at the police precinct level, there is no problem with identifying the location of riots or with obtaining data consistently over time. Of the problems Varshney and Wilkinson identify with government data, NCRB reports solve those related to data availability and the use of a consistent definition of a crime.

However, there has been pervasive criticism regarding the quality of NCRB data on two fronts: the way FIRs are meant to be processed and potential police biases. Whenever an individual comes to a police station to report a riot, a FIR must be completed.[7] Thus, conventional wisdom believes that if several or even hundreds of people come to a police station to report a riot, hundreds of FIRs will be filed, and riot statistics for that police station will

[7] Anju Chaudhary v. State of UP, (2013) 6 SSC 384.

register hundreds of riots as having occurred. This is partially true in that multiple FIRs can be generated for any one riot (Rijiju 2017), but the FIRs must involve different events and individuals.[8] Indian police typically view a riot as consisting of many separate events. For example, if a mob hits an individual during the course of a broader riot, that individual can report the riot to the police, and the police will record a FIR. Anyone else injured in the riot by a different group or in a different area can file a separate FIR. As such, riots in the NCRB data may overestimate the true number of riots. However, NCRB data follows the Principal Offense Rule (POR) wherein only the primary crime is registered in the FIR and included in the NCRB data. This means that some crimes that take place during a riot — for example one-on-one fighting — may primarily fit the definition of assault and secondarily be part of a riot. Thus, while many individuals may report a riot, only a few of these cases are likely to be registered as a riot under the POR (Dubbudu 2015, 1). Bhavnani and Lacina (2015) suggest solving this possible over-counting problem by looking at riot trends over time. Since the FIR and POR system has been in place since independence, over-counting problems are consistent across time.

The bigger complaint against using NCRB data is that not all reported riots end up being recorded in FIRs. Lack of police resources may be partly to blame: if 1,000 people show up to a precinct to file a FIR, the police will be understandably overwhelmed (Rao 2016). Low police capacity has resulted in the introduction of Omnibus FIRs wherein groups of reports are clumped together into a single FIR (Narula 2003, 15). This practice is illegal, as it hinders the ability of police to actually investigate crimes, but the broader problem of lack of resources has persisted for decades throughout India. Along the same lines, people often report that recorded FIRs are

[8] The Criminal Law (Amendment) Bill, 63-C (2013); Lalita Kumari v. Govt. of UP, (2014) 2 SSC 1, sections 39, 66, 111.

incomplete, never investigated, or do not contain information about the accused (Iyer 2002). Low police capacity may again be to blame, though these problems may also stem from police corruption. Still, FIRs were registered and, thus, would be recorded in NCRB data regardless of their quality or whether anyone was prosecuted.

3.2.2. Non-Registration of First Information Reports Has Always Been a Problem

Anecdotal evidence from media outlets suggests that police corruption also results in non-registration of FIRs (e.g., Sonnenberg 2014, 22, 36, fn. 282). If non-registration is systemic, then NCRB data will not accurately show riot trends over time. Police have great incentives not to register FIRs because their performance is evaluated based on the crime rate, which in turn is generated from FIRs (Mehta 2011, 54). Though countless Supreme Court cases have declared that police must register FIRs, there is no legal or criminal penalty for not doing so.[9]

One way to attempt to verify official crime statistics is by correlating them with survey data on individual victimization. The prevailing conclusion is that victimization studies correlate quite well with crime statistics (Gove, Hughes, and Geerken 1985; Klinger 1997; Levitt 1998).[10] Prasad (2013, 47-48) specifically studies NCRB reports and compares them to a victimization

[9] An exception is that improved technology has been shown to decrease non-reporting and, therefore, increase the crime rate over time (O'Brien 2003; Wittebrood and Junger 2002). These findings do not apply to the Indian case since the system for recording FIRs has not changed since independence, despite countless calls dating back decades for changes (Rajasekaran 2013). Boivin and Cordeau (2011) find that local police may undertake concerted efforts to reduce their workload. Such actions can only be maintained at a local police precinct level; therefore, aggregating to the district and state level should address this concern.

[10] Wilkinson (2010, 597) relies on Dreze and Khera (2000) to claim that *Crime in India* reports have unstable trends over time. An examination of Dreze and Khera (2000) shows that the authors use *Crime in India* data and find that it is of high quality.

survey on burglary and theft. He finds that NCRB "records do contain valuable information about actual crime" and that "they can be used to study crime patterns and to test theories."

This makes sense because non-registration has been a long-standing and pervasive problem throughout India since at least 1861 (Rao 2016, 55). Freedom House (2018) finds that people all over India face substantial barriers to getting an FIR completed. Consistently over 50% of Indians, no matter the time or location, mention non-registration as a common practice (Malimath 2003, 106). Sociological studies of police work in India confirm the existence of these problems over many decades (Bayley 1971; Subramanian 2007). Non-registration of crimes is a severe problem throughout the entire world, with perhaps 40% of crimes in Montreal, Canada being not registered (Boivin and Cordeau 2011). Nevertheless, official crime statistics are the preferred method of analyzing crime precisely because any reporting errors persist between regions and through time, thus, preserving time trends.

"[Wilkinson] concludes that important factors influencing variation in how local police record rioting are likely to be enduring traits of particular states, such as levels of police corruption" and states that "[government] statistics provide an accurate picture of overall trends in Hindu-Muslim violence" (Bhavnani and Lacina 2015, 771; Wilkinson 2004, 244). This means that comparing time trends between states and districts is still appropriate because, although the absolute number of riots may be biased up or down in a particular state or district, the trends in each will be accurate (Dreze and Khera 2000). This can be contrasted with the TOI data, where editorial biases are unknown and certainly change over time.

4. Research Design and Methods

We collect new data on NCRB riot reports at the police precinct level from 1971 to 2005 in Bihar, Chhattisgarh, Gujarat, Jharkhand, Karnataka, Madhya Pradesh, Uttar Pradesh, and Uttarakhand. While resource limitations prevented collecting NCRB data for all police precincts in all states, Varshney (2003) finds that these eight states vary widely their number of riots. Location, population, police resources, and demographics all vary considerably, so conclusions from these eight states should be broadly applicable to riots elsewhere in India. We aggregate the police precinct data to the district level so that it can be directly compared with the TOI dataset. Using the locations of riots listed in the TOI dataset, we identify the appropriate district for each riot in these states for the period from 1971 to 1995. It is important to note that Chhattisgarh, Jharkhand, and Uttarakhand were created from Madhya Pradesh, Bihar, and Uttar Pradesh respectively after 1971. Therefore, we aggregate data from these newly created states into their former states such that the final dataset consists of district-year observations from 1971 to 1995.

The TOI data are meant to capture only Hindu-Muslim riots, whereas the NCRB data includes all riots regardless of their type. As such, we may find differences in riot trends between the two datasets because either Hindu-Muslim riots do not reflect overall riot trends or because one of the datasets is missing riot reports contained in the other dataset.

Toward the first point, despite the TOI dataset purporting to only capture Hindu-Muslim riots, scholars tend to use the TOI data as a proxy measure for overall riot trends (e.g., Urdal 2008). These scholars argue that Hindu-Muslim riots broadly follow the same trends as do other types of riots. A long line of research argues that riots begun for economic (Hasan 1982), political (Engineer 1988), and social (Basu 1994) reasons are frequently portrayed as part of a larger communal narrative (Brass 1997; Brubaker 2004; Kalyvas 2003).

Second, we look for riots reported in the TOI that are not reflected in the NCRB dataset. Finding such riots would imply that the NCRB dataset is incomplete and that this may be a reason for diverging trends between the datasets. We examine district-years with fewer than the first quartile of NCRB riot reports (73) and with at least one TOI riot report; there are 17 such instances (0.45% of the dataset). This means that the NCRB dataset tends not to under-report riots found in the TOI dataset. Given these two points, differences in riot trends are likely the result of either an unexpected overabundance of non-Hindu-Muslim riots or the TOI under-reporting Hindu-Muslim riots.

In comparing these datasets, we use two methods: trend analysis and regression analysis. By analyzing trends, we are able to hold constant many differences in the ways in which the TOI and NCRB report riots and see if common shocks (i.e., actual riot events) impact the two sources in similar ways. We conduct a trend analysis at the aggregate, state, and district levels. Scholars have used TOI data at all three of these levels; thus, it is important to determine if the two sources comport across these units of analysis.

Second, we use regression models to explain both TOI and NCRB riot reports. This analysis helps us to determine whether outside factors systematically influence riot reports. We conduct this analysis at the district-year level. We review the results from each of these methods in turn.

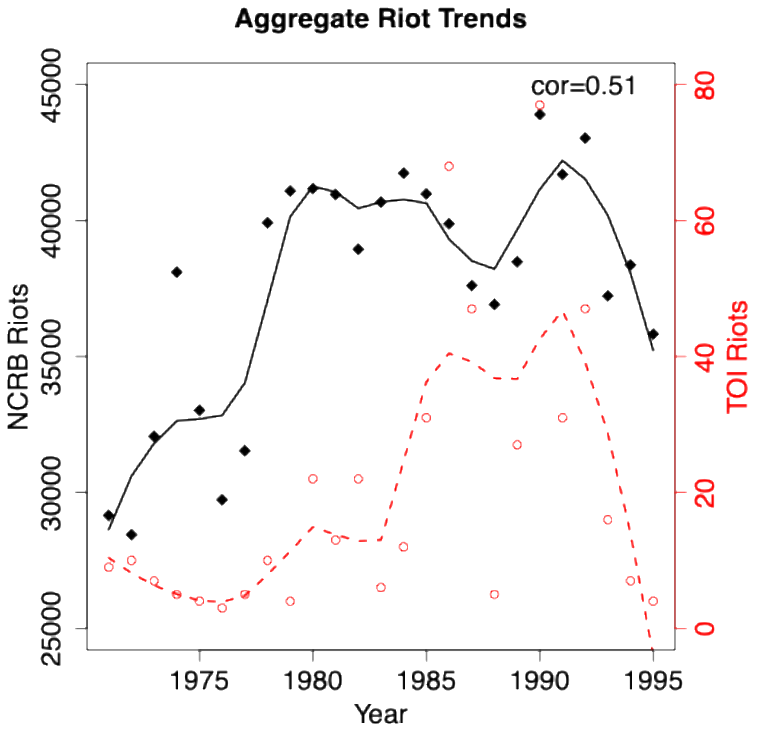
5. Results

5.1. Aggregate Trends

To establish an overall trend in the number and severity of riots over time, we aggregate the NCRB and TOI data to include the states mentioned earlier. These states represent over half the

population of India. In the aggregate sample shown in Figure 1, riot prevalence increases until 1994.[11] The left *y* axis is the number of NCRB riot reports and the right *y* axis is the number of TOI riot reports. The *x* axis is the years during which the TOI and NCRB data overlap. The NCRB data shows a drop in riot activity during the 1980s that the TOI data initially misses, but the general shape is similar. Importantly, NCRB riot reports remain high in the early 1990s while TOI reports drop. Here and throughout the analysis, we compute Pearson correlations to test the strength of the relationship between the TOI and NCRB. The correlation between the TOI and NCRB datasets is a moderately strong 0.51. This finding shows that the TOI and NCRB data display relatively similar trends when aggregated.

Figure 1: Moderate Correlation Between Government Data and Newspaper Data



Note: TOI in circles with dashed line, NCRB in diamonds with solid line. LOWESS smoother with 0.40 span.

[11] The curves in all plots represent LOWESS (Locally Weighted Scatterplot Smoothing) fits. This plot uses a smoothing parameter of 0.40; other plots use 0.25 to account for greater variation year-over-year.

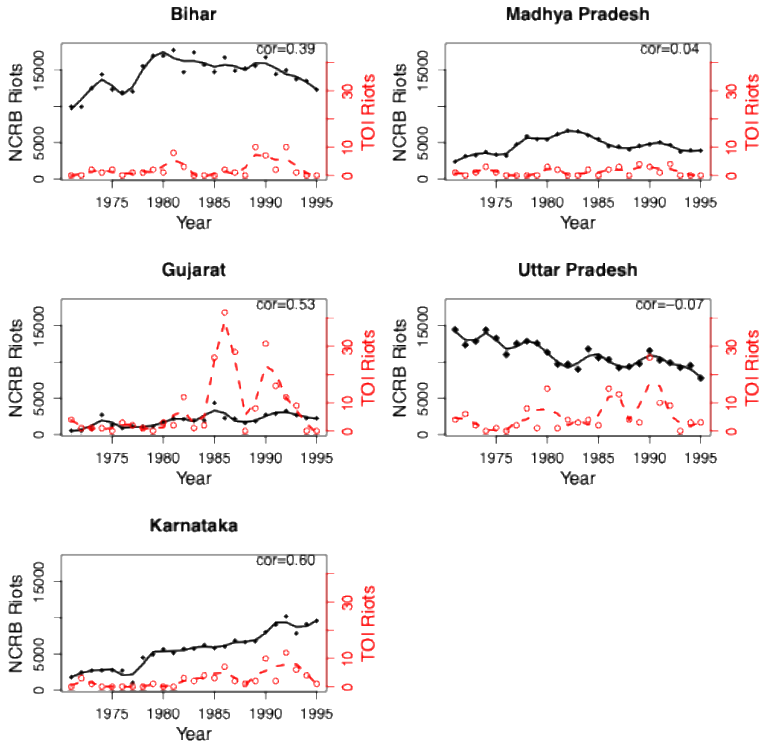
5.2. State Trends

India is a large country with variation across states on almost all possible measures. For that reason, many scholars studying violent collective action know that an aggregate overview of riot trends is insufficient. Sub-units within a country must also confirm aggregate trends in order to develop a broadly applicable theory or to tailor a theory to states that are particularly riot prone.

Varshney (2003) presents the TOI data only at a national level, but Wilkinson (2004) --- Varshney's collaborator in collecting the TOI data --- begins a trend of using these data at the state and district level. Figure 2 shows the data broken out by state. It is immediately obvious that the NCRB data shows many patterns not present in the TOI data. The *y* axis limits for NCRB and TOI data display the minimum and maximum number of riot reports across all states. NCRB and TOI trends are most similar in Gujarat and Karnataka. In Gujarat, riot reporting is relatively low for both datasets until the 1985 Gujarat riots. During this time, the NCRB data shows a small increase in riot reports whereas TOI data spikes to the highest level in any of the states. Media scrutiny of these riots was especially prominent, which might explain why the TOI reports dramatically increased while NCRB reports increased only slightly. Although Karnataka has the strongest correlation between NCRB and TOI data, NCRB riot reports are clearly increasing over time while TOI riot reports are high only around 1990. The correlation in Bihar is very weak. Additionally, NCRB data reports that Bihar is the most riotous state throughout this period, but TOI data suggests that Uttar Pradesh has more riots. There is almost no correlation between the TOI and NCRB data in Madhya Pradesh and Uttar Pradesh. A more important problem emerges in Madhya Pradesh where, even though riot counts for both datasets are relatively stable, the modal number of TOI reports is zero while NCRB reports hover around

5,000 per year. It is implausible that thousands of NCRB riot reports do not translate into even a single significant riot event.

Figure 2: Weak Correlation Between Government and Newspaper Data at the State Level

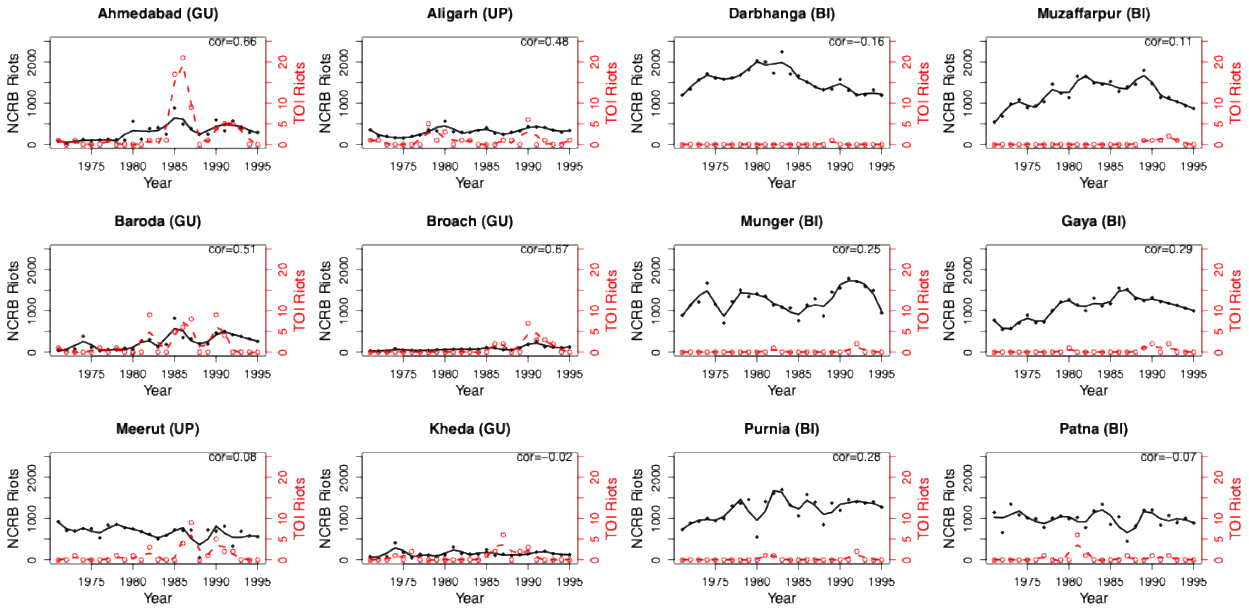


Note: TOI in circles with dashed line, NCRB in diamonds with solid line. Range of plots represents minimum and maximum number of riot reports across five states (as they existed in 1971). LOWESS smoother with 0.25 span.

5.3. District Trends

It is at the district level where the limitations of the TOI data become readily apparent. Figure 3 shows a comparison between the TOI and NCRB in the six districts with the highest number of total riot reports for each dataset. District-years with zero TOI riot reports are common, so choosing districts with many TOI reports biases the results in favor of finding a stronger relationship between the TOI and NCRB. The right panel displays the districts where there were the most NCRB riot reports.

Figure 3: Districts with Largest Number of Newspaper Reported and Police Reported Riots



(a) Districts with Most TOI Reported Riots (b) Districts with Most NCRB Reported Riots

Note: TOI in circles with dashed line, NCRB in diamonds with solid line. Range of plots represents minimum and maximum number of riot reports across all districts. LOWESS smoother with 0.25 span. Left panel is the district with the most TOI reported riots; right panel is the district with the most NCRB reported riots. Plots of the six highest districts on both measures are in the SI.

In Ahmedabad, the TOI data does pick-up the two years with the highest number of NCRB riot reports. Both datasets agree that riots were low in the early and mid 1970s and high in the 1990s, but there is little resemblance between them in the 1980s. However, this comparison is problematic even when the correlation between the TOI and NCRB is high. Ahmedabad experienced 72 TOI riot reports from 1970 to 1995 and 7,270 NCRB reports. This makes Ahmedabad the most riotous district in the TOI data, while on the NCRB measure, Ahmedabad is forty-second. It is apparent that the ratio of NCRB riot reports per TOI report, even in the Ahmedabad case, is very high. The correlations between TOI and NCRB riot reports for the next three most riotous districts in the TOI data are moderate, but there is no relationship between TOI and NCRB riot patterns in Meerut or Kheda.

Darbhanga experiences the most NCRB riot reports, 38,801. At the same time, the TOI reported only one riot occurring from 1971 to 1995. The TOI misses all the variation in riot prevalence over time in this district. The same pattern is present for the other districts shown with almost no TOI riots reported in these districts and no correlation between TOI and NCRB riot reports.

While we acknowledge that some riot events in the NCRB dataset may be part of one larger riot, it is unlikely that 1,900 NCRB reports in these districts did not produce at least one riot event. What is more probable is that the TOI dataset missed a series of riots in many district-years. Perhaps the riots were not covered because the TOI lacked a reporter in the area to devote time to reporting them. Perhaps there were other stories of other riots or other newsworthy events that the TOI needed to inform the public about. Regardless, using the TOI data misses heterogeneity in riot time trends at the district and state level. Evidence from qualitative research in specific district-years --- e.g., Bose (1981) and other event reports in *Economic and Political Weekly* --- confirms that heterogeneity captured by the NCRB and missed by the TOI is not just random noise, but actual riot events.

5.4. Modeling Riot Coverage

We now seek to model TOI and NCRB riot reports to determine whether the two data sources are influenced by the same types of covariates. Our models are at the district-year level. Because the NCRB data can only be collected yearly and is not disaggregated to individual riot events, our analysis is necessarily correlational. Since TOI and NCRB riot reports appear to be correlated, we use NCRB data to explain TOI data and vice versa (*NCRB Riots* and *TOI Riots*).

Additionally, we include a lagged dependent variable (*Lag DV*) to capture temporal dynamics. State-level fixed effects help control for unobserved covariates.

Our other covariates come from the Indian Census. The population (*Log Population*) of a district should partly explain its riot prevalence: districts with more people in them have a greater opportunity for some subset of people to start a riot. The proportion of Hindus compared to Muslims (*Hindu/Muslim*) and the percentage of Scheduled Caste members (*Pct. SC*) describe the causes of common local conflicts. We also include controls for *Population Density* and the amount of agricultural cultivation (*Ag. Cult.*) in a district.

We also include covariates that we expect may be associated with newspaper readership: *Pct. Literate*, *Log Km to Mumbai*, and *Pct. BJP*. The percentage of literate residents in a district captures both the number of potential newspaper readers (Jeffrey 1993) and the extent of the rural/urban divide --- as more rural districts are likely to have high illiteracy. Distance to Mumbai represents the difficulty of TOI newspaper reporters, who are primarily based in Mumbai, accessing districts to report on riots. Percentage of Bharatiya Janata Party (*Pct. BJP*) voters may indicate newspaper partisan biases.

We use negative binomial regression to predict the number of riot reports in the TOI and the NCRB. SI.2 discusses a robustness check using logistic regression. Table 1 displays the results with population as a positive predictor of riot reports across both dependent variables. It makes sense that districts with more population have a greater opportunity for animosity and conflict. Similarly, as the Hindu/Muslim ratio increases, power struggles become less competitive. In line with expectations, fewer riots are reported in the TOI further from Mumbai.

Distance to Mumbai is also statistically significant for NCRB riot reports.[12] While districts with a higher proportion of Scheduled Caste residents report more riots in the NCRB data, a higher proportion of BJP voters in a district is associated with fewer TOI riot reports. As expected, TOI riot reports are also higher in districts with greater literacy.

Table 1: TOI and NCRB Riot Reports

	Dependent variable:	
	TOI	NCRB
	(1)	(2)
Lag DV	0.222***	0.001***
	(0.048)	(0.0001)
NCRB Riots	0.001***	
	(0.0003)	
TOI Riots		0.098***
		(0.012)
Log Population	0.904***	0.595***
	(0.259)	(0.025)
Pct. SC	-1.198	0.946***
	(1.841)	(0.198)
Hindu/Muslim	-0.072***	-0.004***
	(0.016)	(0.0003)
Pct. BJP	-1.995***	-0.024
	(0.515)	(0.066)
Pct. Literate	4.446***	-0.048
	(1.028)	(0.116)
Pop. Density	0.0005	-0.0001
	(0.001)	(0.0001)
Ag. Cult.	-0.158	-0.024
	(0.289)	(0.038)
Log Km. to Mumbai	-0.516**	-0.101***
	(0.257)	(0.035)
Gujarat	1.673***	-0.731***

[12] This is likely an artifact of districts with a history of riots happening to be located closer to Mumbai, not a theory-driven result.

	(0.544)	(0.070)
Karnataka	0.800*	-0.030
	(0.475)	(0.059)
Madhya Pradesh	1.641***	-0.476***
	(0.533)	(0.056)
Uttar Pradesh	0.909**	-0.320***
	(0.410)	(0.051)
Constant	-13.642***	-2.715***
	(4.152)	(0.442)
Observations	2,675	2,671

*p<0.01; **p<0.05; ***p<0.01

Note: Negative binomial regression.

6. Discussion and Conclusion

We contribute to the sociological conversation on the use of event history data by comparing riot reporting across newspaper and official government sources. While both newspaper and government data are frequently used, few studies have been able to systematically compare these two data sources, especially outside of the United States. Further, our location-based data allows us to ascertain how well these two sources compare at different levels of aggregation. Scholars frequently theorize about either national or local collective action and then wish to generalize their results to different units-of-analysis. Our explicit comparison of two data sources across units-of-analysis provides guidance for scholars seeking to make these generalizations. Additionally, we identify factors that influence the prevalence of riot reports in both datasets. This allows us to more precisely describe the potential biases in each data source.

By examining event history data in the context of riots, our results speak both to scholars of political protest and large-scale civil conflict. Riots occur more frequently than does civil conflict, but riots are also more consequential than most other forms of collective action since injury and death often occurs. Event history data has also been used extensively in the study of

riots. In the United States, scholars have used both newspaper and government data in quantitative and qualitative studies of riots, particularly those occurring during the Civil Rights Movement. Outside of the United States, the focus has been on either using the TOI dataset to measure Indian riot events or on case studies of specific riots. We tie these two literatures together with our comparison of newspaper and government data in India. Our case also has practical importance: riots are ongoing in India, and we need to develop better measures of riot events in order to analyze ways to reduce riot prevalence in the future.

We also advance the Indian sociological literature on violent collective action. Prior work has done an excellent job of examining noteworthy riot cases and understanding the reasons why newspapers decide to report on these riots. By comparing TOI and government data, we show district-years where newspapers did not report on riots, but where riots may have occurred. This presents a new approach to studying the relationship between riots and media coverage in India. Future studies would do well to select particularly discrepant cases with high government riot reports and no newspaper riot reports and to investigate the reasons why newspapers and political elites did not publicize these riots. Such an investigation will help fully uncover the ways in which riots are politicized.

Our results provide new guidance for constructing the best measure of collective action events. We make two main recommendations. First, newspaper-based riot measures comport with government data at aggregate levels. This suggests that newspaper data is an appropriate source for information about riot events when scholars are interested in describing broad riot patterns and using these data to identify specific cases of interest. Newspaper data provides a much richer source of information about any given riot, and difficulties about determining the motivations behind any given riot can be addressed through thorough case study analysis.

Second, we find that government riot data is more granular and is less likely to be influenced by factors that impact newspapers' decisions about whether to report a riot. Government data, therefore, is more useful when conducting regression analysis or when examining riots at the sub-national level. Scholars recently have turned to understanding the micro-motivations and local level implications of violent collective action, and government data will be useful in this regard.

We caution that both types of data have limitations that cannot be fully addressed. Adding additional newspapers to the TOI will not improve the accuracy of the dataset because we have no way of controlling for the editorial biases and reporter decisions that go into producing these data. By combining multiple newspaper sources, we actually make it more difficult to discern riot reporting biases because each newspaper operates and reports on riots differently. Scholars can and should compare newspaper-based sources to see whether the same trends in riot events occur over time, but such exercises are more helpful in identifying the biases of different newspaper sources rather than establishing a complete riot event dataset. Government data is limited in its usefulness by the ways in which riot events are coded. In this instance, we cannot precisely identify anything about specific riot events. Our suggestion for constructing the best measure of collective action events is to use and interpret all available data fully aware of the problems associated with each dataset. Not only does this mean replicating results using different strategies for measuring the same types of events, but scholars should think carefully about the ways in which datasets were coded and collected to interpret differences in findings between them.

Developing clear standards for creating and using event data could help scholars to collect event data more systematically across sources and contexts. Doing so would help

consolidate data sources, allow for more comprehensive comparisons between data sources, and identify key areas to focus future data collection efforts. Such an effort could involve scholars, media organizations, non-profit groups, and government support.

With the advent of the Internet, newspapers have become a less important part of life in many countries. Not so in India, where newspaper circulation is still high, and newspapers fill an important social and cultural role. Newspaper event history continues to be relevant in India, but our recommendations regarding the use of event history datasets can be extended and applied to contexts where newspapers play a less dominant role. We do not know what motivates any particular individual to post about a riot on social media. People are likely to post about riots when they have a better Internet connection, but people behave like newspaper editors by publishing content strategically for reasons that are not completely clear. Despite social media providing exponentially more data about riot events, data quality and the mechanisms by which data are produced are concerns now more than ever.

References

Abbott, Andrew. (1995). Sequence Analysis: New Methods for Old Ideas. *Annual Review of Sociology*, *21*(1), 93-113.

Andrews, Kenneth T. and Neal Caren. (2010). Making the News: Movement Organizations, Media Attention, and the Public Agenda. *American Sociological Review*, *75*(6), 841-866.

Barranco, Jose and Dominique Wisler. (1999). Validity and Systematicity of Newspaper Data in Event Analysis. *European Sociological Review*, *15*(3), 301-322.

Basu, Amrita. (1994). When Local Riots Are Not Merely Local: Bringing the State Back in, Bijnor 1988-92. *Economic and Political Weekly*, *29*(40), 2605–2621.

Bayley, David H. (1963). Violent Public Protest in India: 1900—1960. *The Indian Journal of Political Science*, *24*(4), 309–325.

Bayley, David H. (1971). The Police in India. *Economic and Political Weekly*, *6*(45), 2287–2291.

Bhavnani, Rikhil R. and Bethany Lacina. (2015). The Effects of Weather-Induced Migration on Sons of the Soil Riots in India. *World Politics*, *67*(4), 760–794.

Boivin, Remi and Gilbert Cordeau. (2011). Measuring the Impact of Police Discretion on Official Crime Statistics: A Research Note. *Police Quarterly*, *14*(2), 186–203.

Bose, Pradip Kumar. (1981). Social Mobility and Caste Violence: A Study of the Gujarat Riots. *Economic and Political Weekly*, *16*(16), 713–716.

Brass, Paul. (1997). *Theft of an Idol*. Princeton University Press.

Brass, Paul (2003). The Partition of India and Retributive Genocide in Punjab 1946-47: Means, Methods, and Purposes. *Journal of Genocide Research*, *5*(1), 71-101.

Brubaker, Rodgers. (2004). *Ethnicity Without Groups*. Harvard University Press.

Demarest, Lisa, and Arnim Langer. (2022). How Events Enter (or Not) Data Sets: The Pitfalls and Guidelines of Using Newspapers in the Study of Conflict. *Sociological Methods and Research*, *51*(2), 632-666.

Dreze, Jean and Reetika Khera. (2000). Crime, Gender, and Society in India: Insights from Homicide Data. *Population and Development Review*, *26*(2), 335–352.

Dubbudu, Rakesh. (2015). NCRB Contradicts Home Ministry, Says over 1200 Communal Incidents in 2014. https, //factly.in/ncrb-contradicts-home-ministry-over-communal-incidents-says -over-1200-communal-incidents-in-2014/.

Earl, Jennifer, Andrew Martin, John D. McCarthy and Sarah A. Soule. (2004). The Use of Newspaper Data in the Study of Collective Action. *Annual Review of Sociology*, *30*(1), 65–80.

Engineer, Asghar Ali. (1988). Aurangabad Riots: Part of Shiv Sena's Political Strategy. *Economic and Political Weekly*, *23*(24), 1203–1205.

Freedom House. (2018). India. https, //freedomhouse.org/report/freedom-world/2018/india.

Ganguly, Sumit, Larry Diamond and Marc F. Plattner. (2007). *The State of India's Democracy*. Johns Hopkins University Press.

Gove, Walter R., Michael Hughes and Michael Geerken. (1985). Are Uniform Crime Reports a Valid Indicator of the Index of Crimes? An Affirmative Answer with Minor Qualifications. *Criminology*, *23*(3), 451–502.

Hagen, Ryan, Kinga Makovi and Peter Bearman. (2013). The Influence of Political Dynamics on Southern Lynch Mob Formation and Lethality. *Social Forces*, *92*(2), 757-787.

Hasan, Zoya Khaliq. (1982). Communalism and Communal Violence in India. *Social Scientist*, *10*(2), 25–39.

Hoffman, Matthias, Felipe Santos, Christina Neumayer, and Dan Mercea. (2022). Lifting the Veil on the Use of Big Data News Repositories: A Documentation and Critical Discussion of a Protest Event Analysis. *Communication Methods and Measures*, *16*(4), 283-302.

Iyer, Vaidyanathapuram Rama Krishna. (2002). Crime against Humanity: An Inquiry into the Carnage in Gujarat, Findings and Recommendations. Vol. 1. *Citizens for Justice and Peace*.

Jain, Savyasaachi. (2017). Rethinking Media Systems: Insights from a Case Study of Paid News in India. *University of Westminster*.

Jeffrey, Robin. (1993). Indian-Language Newspapers and Why They Grow. *Economic and Political Weekly*, *28*(38), 2004–2011.

Johnson, Erik, Jonathan Schreiner, and Jon Agnone. (2016). The Effect of *New York Times* Even Coding Techniques on Social Movement Analyses of Protest Data. In *Narratives of Identity, Social Movements, Conflicts and Change*, edited by Landon Hancock, 263-291. Emerald Publishing Group.

Kalyvas, Stathis N. (2003). The Ontology of Political Violence: Action and Identity in Civil Wars. *Perspectives on Politics*, *1*(3), 475–494.

Klinger, David A. (1997). Negotiating Order in Patrol Work: An Ecological Theory of Police Response to Deviance. *Criminology*, *35*(2), 277–306.

Lambert, Richard D. (1951). Hindu-Muslim Riots. *University of Pennsylvania*.

Levitt, Steven D. (1998). The Relationship between Crime Reporting and Police: Implications for the Use of Uniform Crime Reports. *Journal of Quantitative Criminology*, *14*(1), 61–81.

Malimath, V.S. (2003). Committee on Reforms of Criminal Justice System. Vol. 1. *Ministry of Home Affairs*.

Marsh, Harry L. (1991). A Comparative Analysis of Crime Coverage in Newspapers in the United States and Other Countries from 1960–1989: A Review of the Literature. *Journal of Criminal Justice, 19*(1), 67–79.

Mehta, Swati. (2011). Police Reform Debates in India: Selected Recommendations from the National Police Commission, Ribeiro Committee, Padmanabhaiah Committee Police Act, Drafting Committee, Supreme Court Directives in Prakash Singh v/s Union of India. *Commonwealth Human Rights Initiative*.

Mody, Bella. (2015). How Well Do India's Multiple Language Dailies Provide Political Knowledge to Citizens of This Electoral Democracy? *Journalism Studies, 16*(5), 734–749.

Murthy, C.S.H.N., Challa Ramakrishna, and Srinivas R. Melkote. (2010). Trends in First Page Priorities of Indian Print Media Reporting-A Content Analysis of Four English Language Newspapers. *Journal of Media and Communication Studies*, *2*(2), 39–53.

Myers, Daniel J. and Beth Caniglia. (2004). All the Rioting That's Fit to Print: Selection Effects in National Newspaper Coverage of Civil Disorders, 1968-1969. *American Sociological Review, 69*(4), 519–543.

Narula, Smita. (2003). Compounding Injustice: The Government's Failure to Redress Massacres in Gujarat. *Human Rights Watch*.

Nayar, Baldev Raj. (1975). Violence and Crime in India: A Quantitative Study. *Macmillan Company of India*.

O'Brien, Robert M. (2003). UCR Violent Crime Rates, 1958–2000: Recorded and Offender-Generated Trends. *Social Science Research, 32*(3), 499–518.

O'Brochta, William. (2019). Pick Your Language: How Riot Reporting Differs Between English and Hindi Newspapers in India. *Asian Journal of Communication, 29*(5), 405-423.

Oliver, Pamela E. and Gregory M. Maney. (2000). Political Processes and Local Newspaper Coverage of Protest Events: From Selection Bias to Triadic Interactions. *American Journal of Sociology, 106*(2), 463-505.

Oliver, Pamela E. and Daniel J. Myers. (1999). How Events Enter the Public Sphere: Conflict, Location, and Sponsorship in Local Newspaper Coverage of Public Events. *American Journal of Sociology, 105*(1), 38-87.

Olzak, Susan and Suzanne Shanahan. (2003). Racial Policy and Racial Conflict in the Urban United States, 1869-1924. *Social Forces, 82*(2), 481-517.

Ortiz, David, Daniel Myers, Eugene Walls and Maria-Elena Diaz. (2005). Where Do We Stand with Newspaper Data? *Mobilization: An International Quarterly, 10*(3), 397–419.

Prasad, Kislaya. (2013). A Comparison of Victim-Reported and Police-Recorded Crime in India. *Economic and Political Weekly, 48*(33), 47–53.

Rajasekaran, R. (2013). NCRB Follows International Practice to Count Crime. https://www.thehindu.com/opinion/op-ed/ncrb-follows-international-practice-to-count-crime/article5135345.ece.

Rao, U.N.B. (2016). A Study on Non-Registration of Crimes: Problems and Solutions. *Ministry of Home Affairs*.

Rao, Ursula. (2010). News as Culture: Journalistic Practices and the Remaking of Indian Leadership Traditions. *Berghahm Books*.

Rijiju, Shri Kiren. (2017). Lok Sabha Unstarred Question No. 6036.

Snyder, David and William R. Kelly. (1977). Conflict Intensity, Media Sensitivity, and the Validity of Newspaper Data. *American Sociological Review, 42*(1), 105-123.

Sobolev, Anton, M. Keith Chen, Jungseock Joo, and Zachary Steinert-Therelkeld. (2020). News and Geolocated Social Media Accurately Measure Protest Size Variation. *American Political Science Review, 114*(4), 1343-1351.

Sonnenberg, Stephan. (2014). When Justice Becomes the Victim: The Quest for Justice after the 2002 Violence in Gujarat. *International Human Rights and Conflict Resolution Clinic*.

Sonwalkar, Prasun. (2002). 'Murdochization' of the Indian Press: From By-Line to Bottom-Line. *Media, Culture & Society*, *24*(6), 821–834.

Sonwalkar, Prasun. (2004). Mediating Otherness: India's English-Language Press and the Northeast. *Contemporary South Asia*, *13*(4), 389–402.

Subramanian, K.S. (2007). Political Violence and the Police in India. *Sage Publications, Inc*.

Urdal, Henrik. (2008). Population, Resources, and Political Violence: A Subnational Study of India, 1956-2002. *Journal of Conflict Resolution, 52*(4), 590-617.

Varshney, Ashutosh. (2003). Ethnic Conflict and Civic Life: Hindus and Muslims in India. *Yale University Press*.

Varshney, Ashutosh and Steven Wilkinson. (1996). Hindu-Muslim Riots 1960-93: New Findings, Possible Remedies. *Rajiv Gandhi Institute for Contemporary Studies*.

Ward, Michael, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. (2013). Comparing GDELT and ICEWS Event Data. *Analysis*, *21*(1), 267-297.

Weidmann, Nils B. (2015). On the Accuracy of Media-Based Conflict Event Data. *Journal of Conflict Resolution, 59*(6), 1129–1149.

Wilkinson, Steven. (2004). Votes and Violence: Electoral Competition and Ethnic Riots in India. *Cambridge University Press*.

Wilkinson, Steven. (2008). Which Group Identities Lead to Most Violence? Evidence from India. In *Order, Conflict, and Violence*, edited by Stathis Kalyvas, Ian Shapiro, and Tarek Masoud, 271-300. Cambridge University Press.

Wilkinson, Steven. (2010). Data and the Study of Indian Politics. In *The Oxford Companion to Politics in India*, edited by Niraja Gopal Jayal and Pratap Bhanu Mehta, 587-601. Oxford University Press.

Wittebrood, Karin and Marianne Junger. (2002). Trends in Violent Crime: A Comparison between Police Statistics and Victimization Surveys. *Social Indicators Research*, *59*(2), 153–173.

Zhang, Han, and Jennifer Pan. (2019). Casm: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media. *Sociological Methodology*, *49*(1), 1-57.

Supplemental Information: Quantifying and Contextualizing Violent Collective Action Event Datasets

The Supplemental Information (SI) contains SI.1, which compares newspaper riot coverage across four newspaper sources, and SI.2, which presents robustness checks. Replication files are available on the Harvard Dataverse.

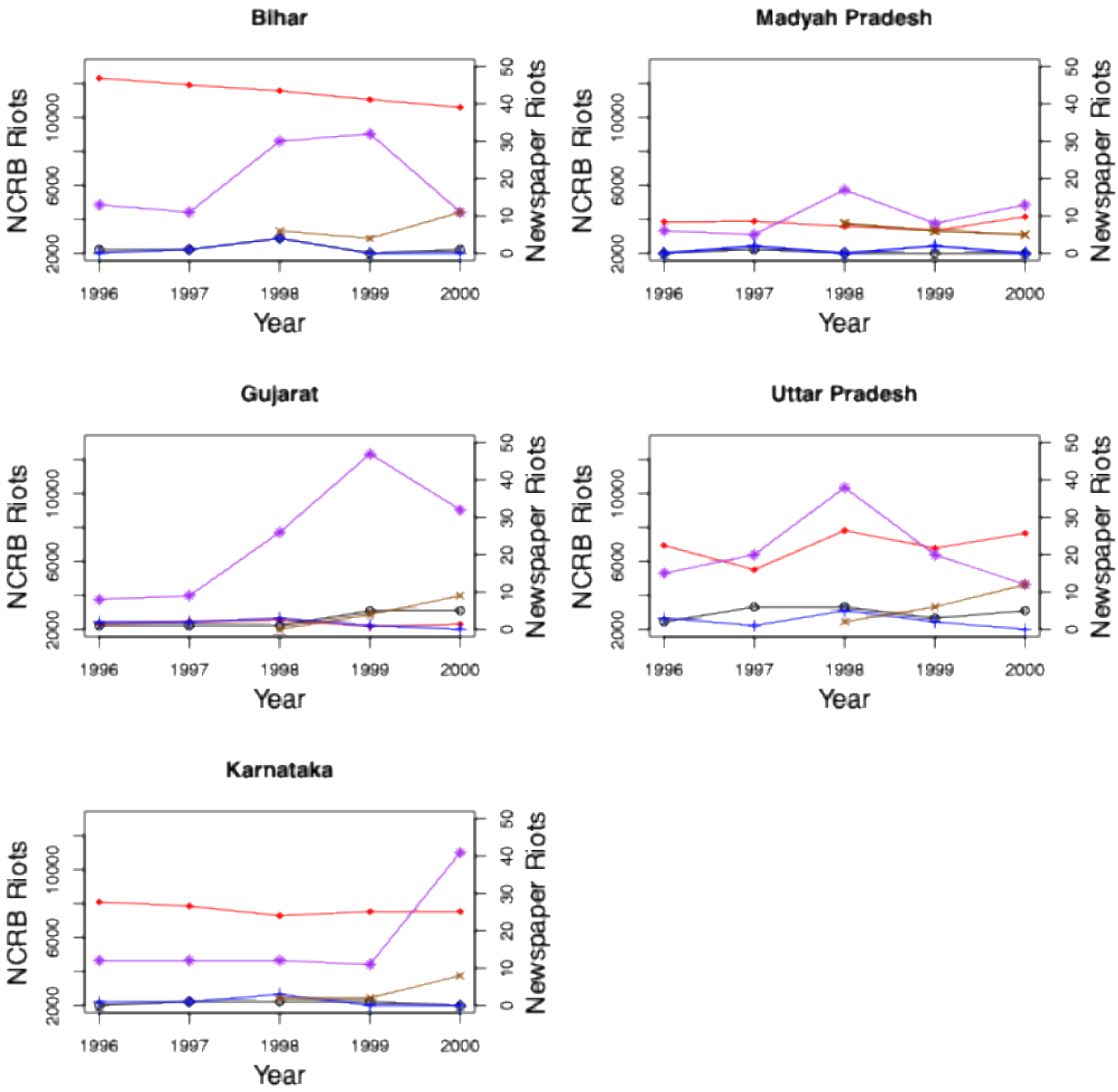
SI.1. Four Indian Newspaper's Riot Coverage

To provide a further confirmation that the TOI newspaper data is non-systematically biased over time, we compare TOI riot reporting with reports of riots from other Indian newspapers. We find that newspaper riot reports show different time trends for the same regions. For this analysis we use a version of the TOI dataset from Anirban Mitra and Debraj Ray that extends the dataset through 2000. We searched all major English and Hindi language newspapers in India for newspaper archives that pre-dated 2000. Only three newspapers resulted: The Hindu (1996-2000), The Tribune (1998-2000), and India Today (1996-2000).

The first author scraped search results for articles in these newspapers with the format "riot" and [state name] for each state in India. This is a crude measure of riot reporting since not all articles mentioning the word "riot" will be about a riot occurring in the year the article was published. However, further data cleaning would introduce additional bias by requiring coding judgments on whether an article was relevant or not.

Figure SI.1.1 shows the trends for the five Indian states where police riot reports are available along with the four newspaper sources. None of the newspapers compare very well with each other, thus adding to our conclusion that newspaper data is non-systematically biased.

Figure SI.1.1: Comparison of Riot Reporting in Four Indian Newspapers



Note: TOI in black circles, NCRB in red diamonds, *The Hindu* in purple stars, *India Today* in blue diamonds, and *The Tribune* in brown crosses.

SI.2. Robustness Check

We present a robustness check to the main analysis using logistic regression. To specify the logistic regression model, we first needed to determine how to count ones and zeros. Out of the 2,675 possible district-years, the TOI reports zero riots occurring in 2,456 or 91.8% of cases.

This contributes to the high ratio of 1,900 NCRB reports for every one TOI report. On the other hand, there are only three district-years (0.11%) with zero riot reports in the NCRB data. To ensure that we use the same empirical specification across models, we define ones and zeros for the NCRB data based on the first quartile of riot reports. Districts with more than the first quartile of riot reports (73) are coded as one, with the remaining districts coded as zero. It is not possible to employ a zero-inflated model or a standard logistic regression model across both datasets because the NCRB data has too few zeros. The model results are shown in Table SI.2.1 and are consistent with those displayed in the main text.

Table SI.2.1: TOI and NCRB Riot Reports

	Dependent variable:	
	TOI	NCRB
	(1)	(2)
Lag DV	1.015***	2.725***
	(0.188)	(0.157)
NCRB Riots	0.936***	
	(0.347)	
TOI Riots		0.638
		(0.389)
Log Population	1.065***	2.071***
	(0.276)	(0.235)
Pct. SC	0.432	5.221***
	(1.895)	(1.593)
Hindu/Muslim	-0.066***	-0.010***
	(0.017)	(0.003)
Pct. BJP	-1.927***	0.131
	(0.556)	(0.478)
Pct. Literate	2.940***	-0.866
	(1.100)	(0.855)
Pop. Density	0.001	-0.0004
	(0.001)	(0.001)
Ag cult	0.080	-0.247

	(0.295)	(0.323)
Log Km. to Mumbai	-0.398	-0.525**
	(0.281)	(0.268)
Gujarat	1.507***	-16.841
	(0.550)	(522.472)
Karnataka	0.674	-14.381
	(0.479)	(522.472)
Madhya Pradesh	1.481***	-16.059
	(0.534)	(522.471)
Uttar Pradesh	0.239	-15.800
	(0.362)	(522.471)
Constant	-17.335***	-10.493
	(4.268)	(522.485)
Observations	2,675	2,671

*p<0.01; **p<0.05; ***p<0.01

Note: Logistic regression.