CONSTRUCTING RELATIONAL AND VERIFIABLE PROTEST EVENT DATA: FOUR CHALLENGES AND SOME SOLUTIONS*

Pamela Oliver, Alex Hanna, and Chaeyoon Lim†

We call for a relational approach to constructing protest event data from news sources to provide tools for detecting and correcting errors and for capturing the relations among events and between events and the texts describing them. We address two problems with most protest event datasets: (1) inconsistencies and errors in identifying events and (2) disconnect between data structures and what is known about how protests and media accounts of protests are produced. Relational data structures can capture the theoretically important structuring of events into campaigns and episodes and media attention cascades and cycles. Relational data structures support richer theorizing about the interplay of protests and their representations in news media discourses. We present preliminary illustrative data about Black protests from these new procedures to demonstrate the value of this approach.

Protest event analysis is important for assessing variations in movement activity and dynamics across time, place, and issue. But there are two serious problems with protest event data. First, they are riddled with inconsistencies and errors that often can be neither detected nor corrected. There is a need for improved verifiability of protest event datasets. The second problem is that protest event data are typically collected and stored in ways that ignore both the relations among events and between events and the texts describing them. Social movement theory recognizes that protests are relational: they are often structured as campaigns and episodes. Further, they usually involve closely linked events such as protest and counterprotest or coordinated protests in different cities on the same day. Similarly, protests described in news sources are not simple reflections or random samples of protests but are selected and filtered through media processes. The volume and style of media coverage influence the meaning and impact of protest events. But most protest event datasets treat events as independent from each other and the texts describing them. Citations to texts are treated as documentation of events but not as data in themselves.

Scholars of social movements have long recognized the mutual relationship between social movements and the news coverage they receive (Gitlin 1980, Vliegenthart and Walgrave 2012). Still, this mutuality has rarely been a quantitative protest event research focus. Instead, most projects have focused on either constructing event catalogs from news sources or studying news coverage of selected protests. Our central methodological innovation is to create relational data structures, which (with the technology that enables them) permit both event-centric and media-centric approaches to be used in the same project and brought into dialog with each other. We treat protests as events involving people in time and physical space and, as the news article mentions, in discursive space. Relational data structures can also capture the structuring of events into campaigns and episodes and permit the aggregation and disaggregation of complex and aggregate events. Finally, relational data structures provide better tools for quality control and verifiability of coding for inherently ambiguous, complex, and relational events.

* Pamela Oliver is Professor Emerita of Sociology at the University of Wisconsin – Madison. Alex Hanna is Director of Research at the Distributed AI Research Institute. Chaeyoon Lim is Professor of Sociology at the University of Wisconsin – Madison. Direct correspondence to pamela.oliver@wisc.edu.

† This research was funded by National Science Foundation grants SES1423784 and SES 1918342. The authors thank Morgan C. Matthews, David Skalinder, and John Lemke for research assistance and conversations that contributed to the development of our methods. Our research protocols and recommendations have developed substantially across the course of this work through several conference presentations and working papers posted to SocArXiv since 2017. The online appendix and replication data are posted at https://osf.io/mp8gs/.

DOI 10.17813/1086-671X-28-1-1

This article introduces our relational approach to protest data collection and discusses how it helps address the key methodological challenges in collecting reliable and valid protest event data that also captures its inherent relationality. After sketching the current state of protest event methods and introducing our project, we discuss how we address four key challenges in protest event data: (1) the problem of errors and verifiability of protest event data; (2) the problem of counting events; (3) the problem of capturing the structuring of events around issues, and (4) the opportunity to treat news coverage as data. We illustrate these issues with examples from our Black protest dataset to demonstrate the power of our relational approach. We conclude with a brief discussion of the implications for future protest event studies.

This project builds on prior work and uses automation to sift through large news media databases to find the one to five percent of articles relevant to protest. Like many other teams, we find that accurately coding events within articles requires human judgment. However, we differ from most other teams in our attention to multiple reports of the same event, often called "duplicates." Many protest event projects have been built around the incorrect assumption of a one-to-one correspondence between articles and events. "Deduplication" is often treated as an afterthought or nuisance. By contrast, we treat repeated mentions of an event as a central feature of texts about events. For example, although 69% of the events in our Black protest data were mentioned in only one article, only 7% of the instances of an article mentioning a protest involved only one event per article and only one article per event. Moreover, 58% of the articles mentioned multiple protests, and 76% mentioned at least one "duplicate" protest that was also mentioned in other articles. Methodologically, this means that constructing protest event records from news articles inherently involves disentangling reports of multiple events in one article and recognizing multiple reports of the same event in different articles. Theoretically, embracing this reality opens the door to important new research about how news media talk about events in relation to each other.

PROTEST EVENT METHODS

Many have recounted the history of protest event studies (e.g., Earl, Martin, McCarthy and Soule 2004, Hanna 2016, Hutter 2014, Olzak 1989b) as it developed from a compilation of event catalogs of labor strikes and political violence (e.g., Snyder and Tilly 1972, Tilly, Tilly and Tilly 1975) to include protests and demonstrations with attention to rigorously defining events (e.g. Olzak 1989a, Olzak 1989b, Tarrow 1988). Extracting protest events from news sources remains an important and active method in social movement studies. Protocols for collecting protest event data from news archives have become somewhat standardized. Many smaller studies report following practices summarized by Hutter (2014), an update of two older sources (Koopmans and Rucht 2002, Rucht and Neidhardt 1998). Researchers based in the United States (e.g., Martin, Rafail and McCarthy 2017, Rafail 2018, Ratliff 2011, Ratliff 2013) report seeking to replicate the codebook and procedures of the Dynamics of Collective Action (DCA) project (Soule, McAdam, McCarthy John, and Olzak 2009), designed in the 1990s.

Coding protest events in news articles is laborious. The DCA project, which read full texts from the *New York Times* microfilm archives, involved four principal investigators and dozens of graduate students funded over a decade by a series of NSF grants at three institutions. Patrick Rafail's (2019) ambitious project coding local newspapers from twenty cities has required years of work. Lorenzini, Kriesi, Makarov, and West (2022) report having thirty-five graduate students work 642 person-days for six months reading and coding 45,680 news stories previously selected by machine preprocessing.

There are a few fully automated data collection efforts in political science and the defense industry for studying conflict events, beginning with the Kansas Event Data System (KEDS, now CEDS 2013) (Bond, Jenkins, Taylor and Schock 1997, Schrodt and Gerner 1994), its derivative GDELT (GDELT 2021), and Lockheed Martin's Integrated Crisis Early Warning System (ICEWS 2022). Automated approaches have become at least as good as humans at the "haystack" task of identifying news articles that contain some mention of protests and at the specific tasks of

recognizing locations and named entities (Hanna 2017). Fully automated approaches typically capture only sparse information about events: usually date, location, named entities such as individuals or organizations, and relevant action verbs. Machines are much less accurate than humans in recognizing whether an article is describing a current or historical event, in recognizing and disentangling descriptions of multiple events in the same article, in recognizing that events may have taken place in multiple locations or in a different location from the publication location, and in identifying the same event in different sources (Althaus, Bajjalieh, Carter, Peyton and Shalmon 2017, Boschee, Natarajan and Weischedel 2013, Leung and Perkins 2021, Schrodt 2012).

Many research teams interested in accurately capturing more detailed characteristics of protests have concluded that the best approaches, for now, are semiautomated or hybrid workflows that automate the preprocessing of articles to filter out irrelevant articles but rely on human coders at the final steps of identifying and coding events within articles. Major examples include Mass Mobilization in Autocracies (Croicu and Weidmann 2015, Hellmeier, Rød and Weidmann 2019, Weidmann and Rød 2019), the Cline Center's SPEED (Nardulli, Althaus and Hayes 2015), the Zurich-based team studying European protests (Lorenzini et al. 2022, Makarov, Lorenzini and Kriesi 2016), Armed Conflict Location and Event Data (ACLED 2022), Count Love (Leung and Perkins 2021) and Crowd Counting Consortium (CCC) (Fisher, Andrews, Caren, Chenoweth, Heaney, Leung, Perkins and Pressman 2019).

Our project has a similar orientation to these hybrid projects and was developed concurrently with them. We differ from most of them in emphasizing relational data structures linking events to multiple articles,[1] and capturing complex and multidimensional relations between events. Below we discuss these contributions in detail, focusing on how we address some of the key challenges in collecting relational data that are rich and more easily verifiable.

ILLUSTRATIVE DATA

Illustrative data for this article are drawn from a larger project on Black protests in the U.S. We selected newswire articles from 1994-2010 from the *New York Times*, Associated Press Worldstream, and Washington Post/Los Angeles Times services as they are archived in the Annotated English Gigaword (AGW) database available from the Linguistic Data Consortium (Napoles, Gormley and Durme 2012). We used protest-relevant and Black/African American keyword search strings[2] to retrieve a large pool of articles from the AGW database. We used MPEDS, an open-source automated system developed by Alex Hanna (2017), to select the subset of articles that were likely to have information about protests. Hanna (2017) reports that classification errors by MPEDS mimic those of human coders in this inherently difficult task. An automated system for identifying locations in news articles from place names was used to further restrict protests to those in the U.S. (Mediacloud 2020).

We identified 1,346 events from 1994-2010 in 1,210 articles yielding 2,682 instances of an article mentioning an event. Most events are canonical protests such as rallies or marches, but we also include press conferences, riots, boycotts, online and petition campaigns, and other actions relevant to protest or the broader Black movement. Although MPEDS does not search for lawsuits, we include them if they are described in the retrieved articles. We classify 1,109 events as Black or pro-Black, 157 as anti-Black or pro-White, and 80 as non-Black. All events are included in this methodological demonstration.

We code the minimum and maximum estimated number of participants from all available information in the news articles, including references such as "filled the council chambers" or "busloads" and other contextual cues. The range of estimates may be very wide if there is little information in the article. Planned events that may not have occurred have minimum size zero. More details about how we coded numbers of participants are given in the methodological appendix available at https://osf.io/mp8gs/.

CHALLENGE 1: ERROR DETECTION AND CORRECTION

There is increasing attention in science to the problem of being able to check, verify, and replicate research findings (Freese and Peterson 2017). The methodological literature in protest event analysis has many discussions about the difficulties in parsing news articles to retrieve events by either humans or machines (e.g., Althaus et al. 2017, Boschee, Natarajan and Weischedel 2013, Lorenzini et al. 2022, Nardulli and Hayes 2011). As Nardulli and Hayes (2011:1) say, news articles about events "can be convoluted." Good journalism often focuses on the background and context of events rather than event details. Event descriptions may be vague or incomplete, jump back and forth between different events in the same article, occur only toward the end of an article, or be intermixed with contextual information about the issue. Parsing event descriptions may require external knowledge. The first author had to consult a map of New York City to determine that an article that seemed to describe police chasing protesters along I-795 to the Holland tunnel had, in fact, shifted without transition from describing the Brooklyn protest to describing the Holland Tunnel protest. Specific actions and issues often do not fit well into pre-planned coding categories. Articles often contain statements that summarize multiple events, what the SPEED team calls "recapitulations" (Nardulli and Hayes 2011), and we call aggregate events, discussed below. A task this difficult cannot be done consistently and with perfect accuracy in a single attempt by a front-line coder, no matter how well trained.

Checks of machine-coded event data have found accuracy rates as low as 20% and rarely higher than 70% (Schrodt 2012, Wang, Kennedy, Lazer, and Ramakrishnan 2016), with both false positives and false negatives being common. The well-funded and highly regarded SPEED project—which trains coders for 70 hours and tests them for accuracy before putting them into production work—reports that tested coders reliably identified 72-85% of all relevant events and accurately coded the information about these events 75-89% of the time (Hayes and Nardulli 2011). Lorenzini et al. (2022) report that their well-trained coders agreed on event identification 60% of the time and recorded Cohen's Kappa scores for event attributes in the range of .45 to .57, a range they consider "fair to good." Whether this is good enough partly depends on the research goals, but we can safely conclude that even well-funded projects that invest heavily in coder training have nontrivial rates of inconsistencies and even errors in coding.

The concept of a protest event is inherently relational and contextual, involving the three dimensions of actor, action, and claim or issue. The same form may be a protest march or a holiday parade, depending on the reason for the event. Whether the issue meets the usual protest definition of promoting or resisting social change can be ambiguous. Many projects would say that an event calling for Black men to repent and care for their families and communities is not a protest, but that was the stated purpose of the Million Man March, arguably the most important Black movement event of the 1990s. Actions by those with institutional power, such as elected politicians, are generally not considered protests. However, a sit-in conducted in Congress by members of the Black Congressional Caucus probably should be treated as a protest event.

Many teams try to obtain greater agreement about whether an event qualifies for coding with extensive definitions and rules, but there are always ambiguous cases. Lorenzini et al. (2022) report dealing with the problem of ambiguity by asking coders to look for a prescribed list of actions (rather than interpreting purposes) but still report only modest intercoder reliability. Additionally, many teams impose minimum event sizes for events to be coded, even though many news articles provide no explicit size estimates. Despite these challenges, most teams report that all these judgments were made in one pass through a news article.

Matching and Deduplication Errors

When the focus is on creating event records only and verification is about documenting that an event happened, multiple citations of the same event ("duplicates") are an annoyance requiring deduplication. The DCA provides one citation per event; Beissinger (2002) retains up to three citations per event. Some teams report telling front-line coders simply not to code events they

recognize as duplicates, including the older DCA project and the more recent Lorenzini et al. (2022) team, who report discarding 27.3% of their articles because the coder recognized a duplicate event and another 5.3% because duplicate events were found in articles coded by different coders. Count Love (Leung and Perkins 2021) reports all URLs for each event, as does CCC (Fisher et al. 2019) for up to 30 URLs.

Matching up events between different texts (deduplication) is even more error prone than parsing texts for events. The DCA team worked post hoc to address the problem of duplicate events,[3] but it is riddled with apparent deduplication errors. A glaring example is the DCA's coding of the Million Man March, called by Louis Farrakhan of the Nation of Islam, which had at least 400,000 participants on October 16, 1995, and smaller pre- and postmarch events on October 15 and 17. The DCA records for events 9510037 (the 16th), 9510033 (the 15th), and 9510036 (the 17th) all have codes that reference the big rally. The October 16 and 17 size estimates are both 400,000; October 15 lacks a size estimate but has "a million black men" in the "who" field. Neither Louis Farrakhan nor the Nation of Islam appears in any of these records. Event 9512044 records a phantom march of size 100 on October 16 in New York that was created from a cited December article about men getting on buses in New York in October to attend the DC march and follow-up volunteer activities by several groups, including the one named One Hundred Black Men. Another example is event 9401023, whose cited article unambiguously describes two different marches to New York's City Hall by two different groups on the same day that have been jumbled into one record.

These errors in the DCA are readily detected with reference to the cited full texts. Still, some smaller duplicate events are difficult to recognize even in full texts because the different sources describe them differently. For example, different articles described protests about the not-guilty verdict given to the police who killed Amadou Diallo. Protests occurred "at the site of the killing" and "at the victim's home," so matching these descriptions required the contextual knowledge that Diallo was killed on his front steps. Vague or incomplete event descriptions are difficult to de-duplicate. Complex episodes of contention involving flurries of protests in the same city within a few days are challenging to sort out both within and between articles and may require constructing a timeline of events. The assumption that coders can accurately recognize duplicate event reports as they read articles one at a time seems unreasonable. Count Love and CCC imply in their reports that matching events up between sources is unproblematic (Fisher et al. 2019, Leung and Perkins 2021) but offer no direct evidence that they have checked for deduplication errors. Planning for the duplication problem can help, e.g., by preprocessing articles by semantic similarity so that those potentially describing duplicate events can be reviewed together (Boschee, Natarajan, and Weischedel 2013, Leung and Perkins 2021).

Advocates of fully automated systems argue that imperfect data are better than no data and that humans cannot produce real-time protest event data to serve the needs of monitoring the world for emerging conflicts (e.g., Schrodt and Van Brackle 2013). In warning against "data fundamentalism," Charles Taylor (2013:23-24) argues that data should be suitable for the question being asked. However, even high-level approximations can be distorted by failing to recognize duplicate reports of the same event. Mona Chalabi (2014) had to retract an analysis on the high-profile FiveThirtyEight blog about a rise in kidnapping in Nigeria based on the automated GDELT system, which was counting the same Boko Haram kidnapping hundreds of times a day over multiple days (Caren 2014). As we show below, a few events in our data have many duplicate reports.

Relational Data Linking Events and Source Texts

To address these issues, we developed a multistep coding protocol and customized tools to construct a relational database that maintains rigorous links between coded events and all news sources mentioning the event. We code in multiple passes, consulting the source articles as needed to correct errors and inconsistencies. The core of the relational database is a table of articles, a table of events, and a table of article-event pairs linking each event with the articles that cite it.

Figure 1. Database Entity Relations

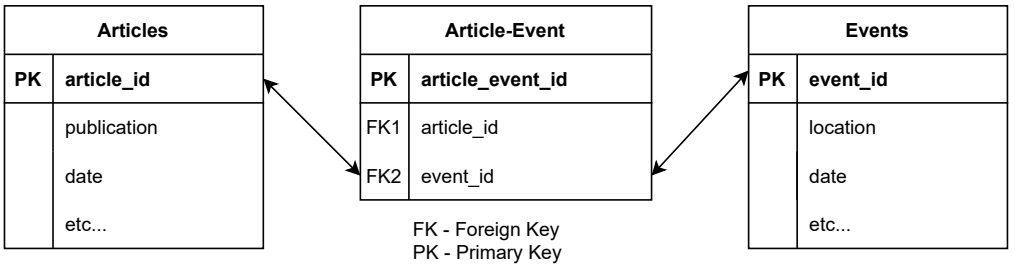
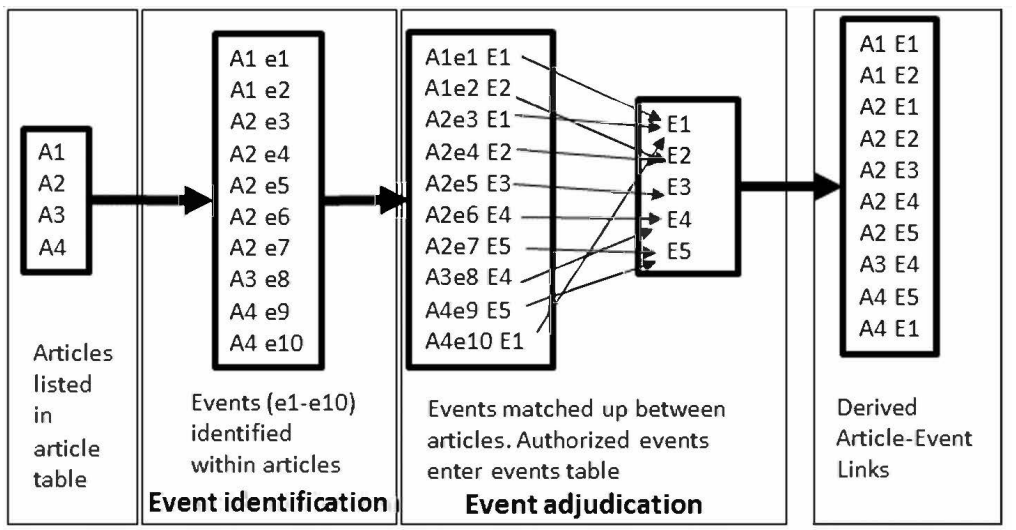


Figure 1 is a database diagram of our core data structure. There is one row in the Article Event table (the middle panel in figure 1) for each instance of an article describing an event, with each unique pair appearing only once. These core tables provide controlled and verifiable links between events and the articles describing them.

Figure 2. Coding Workflow to Match Events between Articles and Create Article-Event Links



This database is constructed in a workflow sketched in figure 2. In the first stage of coding that we call "event identification" phase (the second panel of figure 2), human coders read news articles screened by MPEDS to identify all events in articles, answer questions about the events, and highlight text about them. Coders perform these tasks using Hanna's MPEDS Annotation Interface (MAI) (Hanna 2022), an open-source tool specifically designed for protest event research that we modified to meet our project's needs. Output from the first stage is a table of coder annotations that is fed to the second stage we refer to as "event adjudication" (see the third panel in figure 2). In this step, coders match up events between articles and generate both the authorized event table and the article-event links table (the third and the fourth panels in figure 2). Articles and events have a many-to-many relationship as an article can mention multiple events and an event can be mentioned in multiple articles. Every event is rigorously linked to the articles that mention it and to the text marked by the coders in the event identification phase, which allows researchers to easily revisit the source articles to verify, modify, and augment the coded data. Explicit relationship generation and rigorous maintenance of the article-event table are critical for improving verifiability and replicability of protest data. Every article and every coder annotation are accounted for either by being associated with one or more events or being explicitly ruled

irrelevant. The online appendix provides more detail about our coding procedures and interfaces, including screenshots and operationalizations of key variables.

Linking events and the texts describing them permits verifiability, but it took the development of specialized tools and an explicit review process to identify and correct coding errors. We have done event adjudication twice for these newswire stories. The first time, the first author matched events using spreadsheets that included articles' full texts. The second time, all adjudication was reviewed with a specialized interface built in Microsoft Access by an experienced database programmer who worked closely with the research team and responded to coder feedback. This interface provides specialized forms to give coders the information and tools needed to find and correct numerous mistakes made the first time. Liberal use of free-text description fields and human-readable event names also improve accuracy and permit error detection. Building on our experiences with event adjudication using the Access database, Hanna (2022) has developed an event-adjudication tool that interacts directly with her MAI package.

Although this process adds work and time to treat "deduplication" and error correction as a serious part of the research process, it is the only way to have reliable and verifiable deduplicated event data. We have also redesigned the project workflow: instead of a one-and-done coding of all possible information in the first reading of an article, our project works iteratively, returning to the data to code or recode variables in light of preliminary results. When analysis reveals data errors or inconsistencies, we return to the source files, correct the error, and then regenerate data files for analysis.

CHALLENGE 2: COUNTING PROTEST EVENTS

Buried in protest event analysis is the deep theoretical and methodological question of what is being counted. Most protest event datasets are constructed on the implicit but false assumption that events are independent of and comparable to each other. Table 1 on the following page displays our categorization of the events we have identified. The usual assumption is that a protest is a single physical gathering in time and space: table 1 shows that only 53% of the events we identified fit this simple assumption. Only for this simple case is counting events and participants unproblematic. However, the other event types are all known to be important parts of social movements and protest campaigns.

Complex Gatherings

Protests with counterprotests, large peaceful demonstrations with a small disruptive splinter, multiday sit-ins, or coordinated protests in multiple cities on the same day are simultaneously a single event in one sense but multiple events in a different sense. A total of 21% of the events in table 1 were one-day one-place gatherings that were parts of larger complex events. Four percent are multiday physical gatherings that might have been parsed into a separate event for each day. Deciding whether and how to parse these complex gatherings into distinct events is one of the difficult parts of coding protest events, and there are no universally agreed-upon rules.

A relational approach allows us to record these connections. Operationally, we use the term umbrella for a complex event that comprises multiple events. We create records of event-umbrella links and records associating all the umbrellas with their descriptions and types. This allows us to record the complexity of these events so that appropriate decisions about counting events can be made at the point of analysis. Umbrellas are tagged according to whether they are in one place or multiple places and one day or multiple days. For one-day one-place events, we instruct coders to create distinct events when identifiably different groups of people do different things. So, a hundred people first marching and then sitting-in would be one event, but a large rally of a thousand people followed by ten people sitting in would be two events that are linked via an umbrella. For multiday events, we record one event but provide the information that could expand it to multiple daily events.

Table 1. Number of Events by Issue Groups and Event Types

	Frequency	*Percent* [a]
Physical Gathering One Day One Place		
Stand-alone event	716	53
Part of a one-day one-place complex event	219	16
Part of a one-day multiplace event	52	4
Part of a multiday/place event	15	1
Threatened, may not have happened	9	1
Total	1,011	75
Multiday Physical Gatherings		
One place	44	3
Multiplace	6	0
Total	50	4
Aggregates of Physical Gatherings (only record in data)		
One day one place	22	2
Multiday one place	43	3
One day multiplace	6	0
Multiday multiplace	9	1
Total	80	6
Aggregates of Physical Gatherings (includes other events in data)		
One day one place	1	0
Multiday one place	28	2
One day multiplace	9	1
Multiday multiplace	10	1
Total	48	4
Diffuse Actions		
Diffuse action	25	2
Total	25	2
Strikes and School Boycotts		
Strike	12	1
Total	12	1
Events that are Verbal only		
One day (e.g. lawsuits, statements)	27	2
Multiday (e.g. petitions, online actions, verbal complaints)	51	4
Create organization	8	1
Total	86	6
Consumer Boycotts (nonactions)		
One day	8	1
Multiday	26	2
Total	34	3
Total	1,346	100

[a] Subject to rounding errors

Aggregate References to Physical Gatherings

News stories often refer to protest events in the aggregate with phrases such as "protests last month," "protests in twenty-five cities," or "protests all around the city last night." Reports of "urban unrest" or "riots" also aggregate different actions by different groups of people in different locations. Six percent of our recorded events are aggregate references to events we have no other record of in the newswires (table 1). Even when vague, these references are clues that something happened, which can often be confirmed in other news sources. Another four percent of our events are aggregates for which some but not all the referents are events in our data. Reports of multicity coordinated protests typically state that there were protests in, for example, twenty-five cities, but name only a few cities. We create separate events for each named city plus an aggregate for the unnamed cities and link them all via an umbrella. Other aggregate references are summary recapitulations of events that were reported in more detail in the past. Often these become tag phrases like "four days of rioting" or "daily protests last spring" that anchor subsequent reporting about an ongoing episode of contention and are important for understanding media discourses about protests. Umbrellas link these aggregates to the original events. In all cases, aggregate events are tagged so they can be included or excluded from analysis as appropriate.

Events That Are Not Gatherings

There are other categories of events that are commonly relevant to protest and social movements but are not physical gatherings. Two percent of our events are "diffuse actions," which report that people are organizing, registering voters, or doing civic volunteering in dispersed locations as part of a concerted collective action. One percent are strikes, which involve people not gathering, not being where they would normally be. Strikes are typically accompanied by pickets or protests that may or may not be separately reported. We categorize school boycotts as strikes because students, like workers on strike, are visibly and countably absent. This contrasts with the three percent that are consumer boycotts, for which there is no meaningful direct way to count the numbers of people not buying something. However, we can count the verbal actions of calling for or supporting a consumer boycott. Six percent of our events are verbal only, including lawsuits, issuing statements, concerted campaigns of postcard or telephone complaints, petitions, online actions, and the creation of organizations.

CHALLENGE 3: CAPTURING THE STRUCTURING OF EVENTS INTO CAMPAIGNS AND EPISODES

Scholars of social movements have long recognized the importance of relations among events. Protests tend to occur in cycles or waves (Andrews and Biggs 2006, Tarrow 1989, Wada 2004). Although some of these cycles or waves arise from diffusion processes or external conditions, many arise from intentional campaigns by movement activists (della Porta and Andretta 2002:59), such as the iconic Albany and Birmingham campaigns in the civil rights movement (Morris 1984). Campaigns are often parts of episodes of contention, sequences of action and reaction by both or all "sides" of an issue (Franzosi 1999, Kriesi, Hutter and Bojar 2019, McAdam, Tarrow and Tilly 2001). Protesters may explicitly link their protest to prior protests as part of a campaign or to evoke a politics of memory (Chang and Lee 2021). The claims-making tradition examines the interactions between the claims by different actors as they play out in news sources (Giugni, Koopmans, Passy and Statham 2005, Koopmans and Statham 1999a, Koopmans and Statham 1999b, Koopmans and Olzak 2004). Despite the longstanding recognition by social movement scholars that campaigns or episodes often structure protest

events, most quantitative approaches to constructing protest event data fail to make this structure explicit. Exceptions are Wada (2004) and Kriesi, Hutter, and Bojar (2019), who conduct episode-level analyses.

Specific-Issue Clusters

We capture this structure by attending to specific issues, the things that news articles say protest events are about. Examples of specific issues include discrimination by Denny's restaurant, whether to remove the Confederate flag from the South Carolina capitol, the killing of Amadou Diallo by New York police officers, a proposed antibegging ordinance in the historic civil rights area of Atlanta, celebrating the Martin Luther King holiday, and protesting the inauguration of George W. Bush. We record the relationship between events concerning the same issue by linking events to specific-issue clusters. The term "specific-issue cluster" is operational, not theoretical, and is meant to capture linkages that can include campaigns, episodes, and other types of issues, such as King Day observances. In our relational database, each event has a free-text specific-issue field filled by coders. In the event adjudication phase, we group events about the same specific issue into a specific-issue cluster by creating records in an event-cluster table in the relational database. Coders describe the cluster's specific issue and its location or time boundaries as appropriate and link events to issue clusters. Because news articles typically describe events in relation to other events about the same issue, it is usually straightforward to make these connections while identifying events in articles. A few less straightforward cases require researcher decisions that are documented in the descriptions in the cluster table and revisited at the point of analysis. These include the handling of events organized around abstract issues, coalitional events with long lists of issues, events centered on annual or quadrennial events like King Day celebrations or inaugurations, and nested or overlapping issues. We give examples of how we handled such cases in the online methodological appendix. The general principle is to cluster events that are discussed together in news articles as having a common theme. The same event can logically involve multiple issue clusters; forty-seven events in our data were tied to two clusters and two to three clusters.[4]

Of the 1,346 events we identified, all but 212 (16%) were part of at least one cluster of two or more events about the same specific issue; most of the nonclustered events were described in relation to other nonprotest events that are not in our protest event dataset, such as policy proposals or official actions. The issue clusters vary markedly in size, both in terms of the number of events and the number of participants. Nearly half (49%) of the specific issues are represented by only one event[5]; another 41% involved between two and six events.

A few issue clusters are large. We identified 66 protest events about the 1999 killing of Amadou Diallo by New York police and 50 events around the 2001 killing of Timothy Thomas by Cincinnati police and a related lawsuit; these two clusters alone (0.45% of specific issues) accounted for 8.6% of all events. Another nine clusters with 19 to 34 events accounted for another 16.5%, for a total of 25% of identified events being tied to 2.5% of the issue clusters. These eventful clusters were about the Confederate flag at the South Carolina capitol (34 events); defense of the "Jena 6," six Black teens overcharged after a fight that began with a noose-hanging incident (32); opposing the death penalty for Mumia Abu-Jamal (29), street vendor grievances and other events linked to an arson-murder at Freddy's clothing store in Harlem (24), the New York police killing of Sean Bell (22); the 2000 presidential election (21); the abolition of affirmative action in California (21); a series of fights between Black and Hispanic people in Los Angeles jails (20); and noose-hanging incidents in the fall of 2007, after the Jena 6 mobilization (19).

Events that are part of the same issue cluster are not independent of each other, calling into question the assumptions of many quantitative approaches to analyzing protest event data. However, recognizing the structuring of events into specific-issue clusters lines up with theoretical understandings of how protests are produced and opens the door to new research questions, especially about how news media respond to and feed into these issue clusters.

CHALLENGE 4: PROTEST DATA ARE MEDIA DATA

We know from past research that protest events extracted from news sources are filtered through media-selection processes. No news source covers everything or a random sample of everything; different sources cover different things. The differences among sources are tied to their locations, language, and political and editorial proclivities. The likelihood that any given source reports an event also depends on time-varying factors, including what else is happening at the same time and its sensitivity to different kinds of news (for comprehensive reviews, see Almeida and Lichbach 2003, Earl et al. 2004., Jenkins and Maher 2016, Ortiz, Myers, Walls and Diaz 2005).

The naïve response to the selection issue has been to construct a protest event dataset from an available source or use an existing dataset like the DCA, give some reason why it is a good source, then ignore the selection issue entirely, treating events in the data as equivalent to events on the ground. Articles in this vein are still published regularly. The claims-making tradition similarly treats news sources as unproblematic, although interpreting news media accounts of events as claims in a discursive space defined by the news sources (Koopmans and Statham 1999a, Koopmans and Statham 1999b, Koopmans and Olzak 2004). Amenta and collaborators (Amenta, Caren, Olasky and Stobaugh 2009, Amenta, Elliott and Tierney 2016) also stay entirely within the media framework and study instances of the mention of specific social movement organizations in specific news sources.

Another response has been to seek to expand the pool of sources. Rafail (2019), for example, codes local newspapers from twenty cities. In the extreme, projects like those operated by Cline Center or RAND Corporation scrape the web for millions of news stories. Count Love searches more than 3,000 local news sources that include local newspapers, radio stations, and television stations (Leung and Perkins 2021). CCC further expands the pool to social media and movement organization websites (Fisher et al. 2019). But even such expansive efforts have been criticized for using only English-language sources with their attendant biases (Herkenrath and Knoll 2011) and for adding even more uncertainty about what is or is not being covered in the news sources (Jenkins and Maher 2016). The ACLED project takes the opposite approach and reports that it emphasizes local sources and develops a different set of sources for each country or even region within a country, testing each source for credibility (ACLED 2020). Many who have addressed the problem have concluded that there are no perfect or comprehensive sources and argue that all media and official sources should be treated relatively and compared with each other. (e.g., Jenkins and Maher 2016, Maney and Oliver 2001, Strawn 2008).

The events-only and media-only traditions miss the opportunity to examine the interplay between events and their media coverage. Vliegenthart and Walgrave (2012) note that it is odd that so much social movement scholarship has treated the selectivity of media coverage as a methodological problem rather than a theoretically interesting topic. Social movements need and want media coverage, and the field supports a cottage industry in movement communication strategies. As Vliegenthart and Walgrave (2012) show in their comprehensive review, media coverage affects movements (e.g., Andrews and Biggs 2006, Gitlin 1980, Myers 2000), and movement actions affect media coverage (e.g., Cancian and Ross 1981, Rafail, McCarthy and Sullivan 2019). This positive feedback can explain media cascades (Seguin 2016) in which a few events come to be mentioned repeatedly (Andrews and Caren 2010, Seguin 2016, van de Rijt, Shor, Ward and Skiena 2013, Walgrave and Vliegenthart 2010).

Recording the relations between news articles and the events they report opens the door to new studies that move beyond the "selection bias" paradigm to examine the mutual effects of media coverage and events. Much research addresses variations in news media descriptions of protests (e.g., Campbell, Chidester, Royer, and Bell 2004; Martin, Rafail and McCarthy 2017; Weiner 2011). Reporters do not just list random bits of information about events but rather construct narratives both within articles, and across articles (Davenport 2010).

Acknowledging directly that protest event data are data about media actors and actions turns a methodological conundrum into an arena for research. Which events are brought into relationship with each other within articles? How do different media sources vary in this? Which events are described repeatedly, and which are barely mentioned? How is the news coverage of protest events structured? Does that structure affect how protests have an impact?

Explicitly Linking Media Reports to Events

Our relational approach allows researchers to address these questions by providing data about variations in the intensity of news coverage of events and thus their impacts on public discourse. We identified 1,346 events in 1,210 newswire articles and a total of 2,682 instances of an article mentioning an event (i.e., article-event pairs). Although 69% of the events were mentioned only once, four events were each mentioned in more than forty news articles. The fourteen most frequently mentioned events (described in fourteen or more articles, 1% of all events) accounted for 17% of the news mentions of events; the 16% of the events described in three or more articles accounted for half of the news mentions of events.

Events were typically discussed in relation to other events, usually ones in the specific-issue cluster. Of the 2,682 article-event pairs, only 177 (7%) involved one article per event and one event per article. Of the 1,210 articles, 58% (700) mentioned more than one event, and 76% mentioned at least one event that was also mentioned in other articles. Of the 510 articles that mentioned only one event, for 65% of them the event was a duplicate also mentioned in other articles.

Issue Clusters and Media Attention

The discursive space in the newswires was highly focused on a few specific issues. There was only one newswire article about 61% of the specific-issue clusters, accounting for 22% of the articles. In contrast, five of the 440 clusters (1%) were mentioned in 60 or more articles, accounting for 28% of the articles. These were the protests around the 1999 New York police killing of Amadou Diallo (76 articles), the "riot" and protests and lawsuit around the 2001 Cincinnati police killing of Timothy Thomas (73), the 2007 online mobilization and protests around the Jena 6 (66), protests in 2000 about whether to remove the Confederate flag from the South Carolina capitol (62) and the 1995 Million Man March (60). Another eight clusters were discussed in twenty to forty-two articles, giving a total of 46% of the articles focused on these thirteen issues (3%). Our empirical work (Oliver, Lim, Matthews, and Hanna 2022) gives qualitative attention to these "top" issues and how they were discussed.

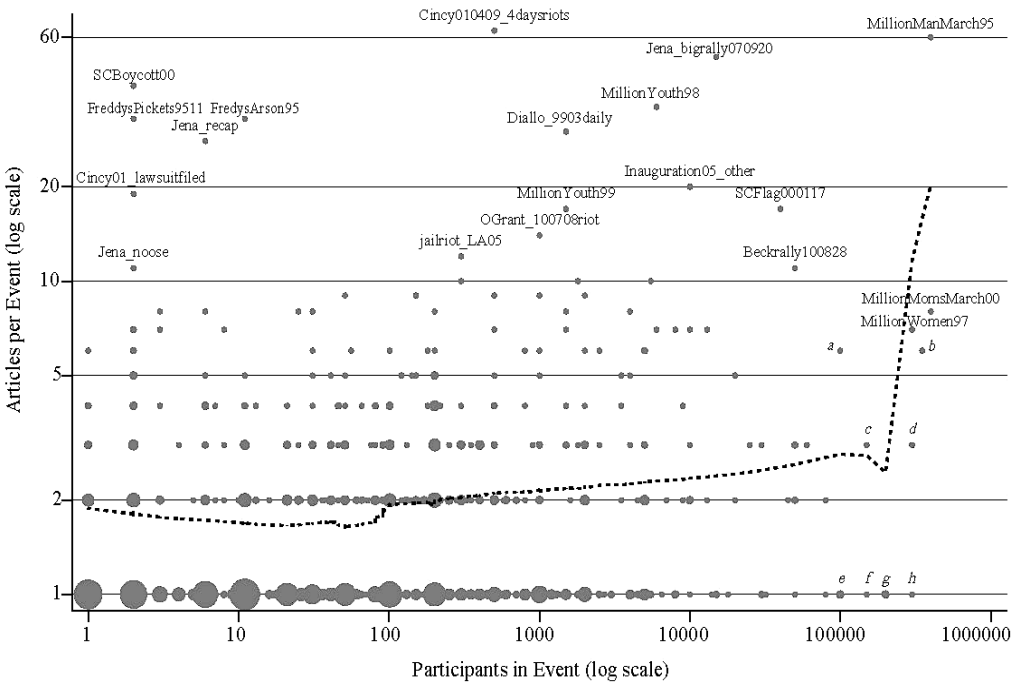
Our data structure allows us to ask whether attention to these top specific issues affected coverage of related issues. The 2007 Jena 6 protests, often called the "new civil rights movement" at the time, began with a noose-hanging incident in 2006. Our data show a flurry of protests about noose-hangings in the months after the Jena 6 protests, but few noose stories otherwise. Similarly, stories about protests against Confederate symbols tend to cluster around the 2000 South Carolina protests. Protests about police killings occur throughout the data, but a more detailed analysis could explore whether the highly covered episodes around police killings lead to more newswire attention to police killings in other places.

Events, Articles, and Participants

Michael Biggs (2016) argues that it is important to attend to the number of participants in protests, not just the number of protests. However, the discursive impact of events is not a simple function of their number of participants. As seen in figure 3, the number of articles about an event is weakly correlated with the minimum estimated number of participants ($r = .223$, $r^2 = .050$); the correlation increases only slightly if the 157 events that are not physical gatherings are excluded ($r = .256$, $r^2 = .066$). The huge Million Man March got a great deal of news coverage, but some very large

events received little coverage.[6] The rise in news coverage by size is modest for most of the range of event sizes and is even weakly negative for events with fewer than 100 participants. The South Carolina tourism boycott had a big impact and received frequent mention but assigning participants to it was problematic. Some very small events received a great deal of news coverage, especially the handful of daily pickets by Black street vendors that preceded the 1995 arson/murder in Harlem's Freddy's clothing store, the hanging of nooses on a tree at Jena high school, and subsequent fights that precipitated the mobilization for the large Jena 6 marches, and the 2001 Cincinnati lawsuit about police discrimination. These small events received a great deal of coverage because they were related to other events, a point we develop below. Most of the events that received extensive coverage (more than ten articles) had thousands of participants, but not hundreds of thousands.

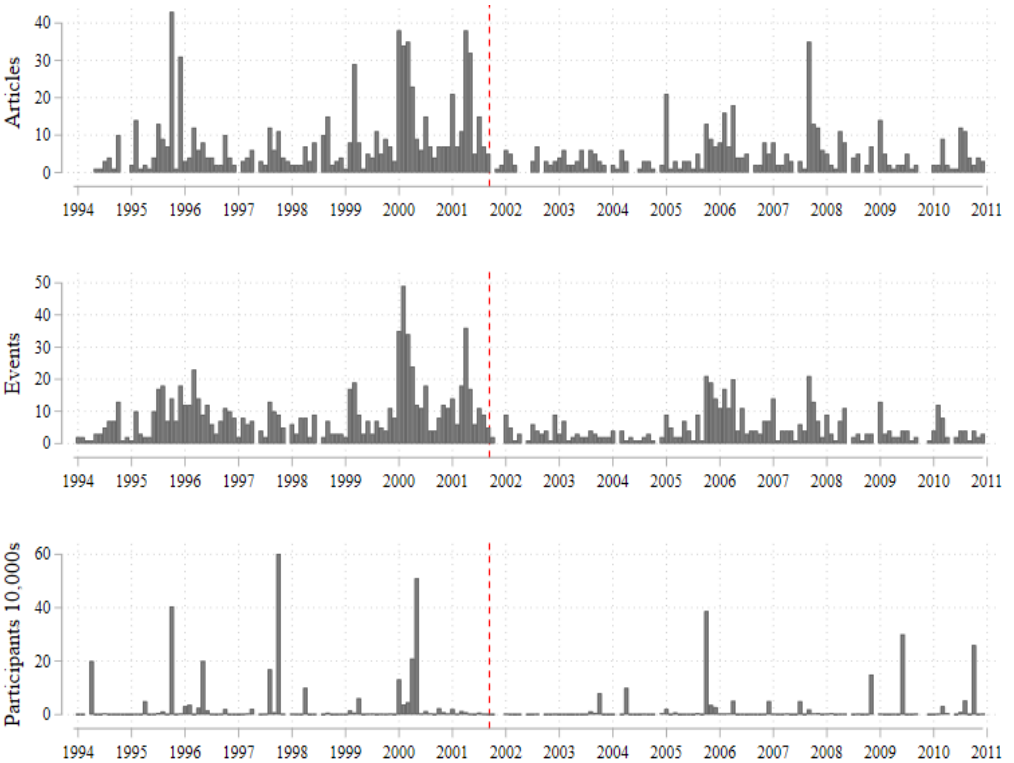
Figure 3. Articles per Event by Participants in Event



Notes: [a] Obama081104; [b] MillionsMore05_rally; [c] OneNationRally1010; [d] PromiseKeepers97; [e] DaytonaBlack-Reunion98, MarchWomensLives04, and MillionMomsMarch00_multi; [f] UPS97_strike; [g] Freaknik94, Millenium-March00 and StandForChildren96; [h] GlennBeck_boycott. Marker size is proportional to frequency. 69% of events had only one article and 95% less than five articles

Figure 4 on the next page plots the monthly counts of protests, articles about protests, and estimated numbers of protest. A vertical line marks 9/11/2001 as the terror attacks appear to have led to a reduction in both Black protests and articles about them, a reduction attested in qualitative accounts of the period (Fletcher and Rogers 2014, Taylor 2016). Article counts track events more than participants, with participation occurring in a few sharp spikes. Visually, the peaks and valleys of protests and articles about protest are similar but not identical. The correlation between articles per month and events per month is high but not perfect ($r^2 = .63$), while the correlation between the number of participants per month and the number of articles per month is low ($r^2=.07$).[7] These results clearly show the importance of tracking the number of events, the number of articles about the events, and the number of participants separately, which our relational approach makes possible.

Figure 4. Monthly Counts of Articles, Events, and Participants



Events, Participants, and Issue Coverage

Our methods allow us to ask which specific issues received more news coverage and what makes some specific issues more newsworthy than others. As with the monthly counts, the number of articles about an issue is more strongly correlated with the number of events (r^2=.73) than the number of participants (r^2=.07) or the log of the number of participants (r^2=.14). Various multivariate specifications all yield the conclusion that number of events is the main predictor of number of articles about a specific issue, with the number of participants adding about .03 to the total R^2. Moving back to the predictors of the number of articles mentioning any given event, we find that the number of events in an event's issue cluster has an independent positive effect on the number of articles mentioning an event. Whether the event's size or the number of events in the cluster is the stronger predictor depends on how the model is specified, but in all specifications, events linked to more other events via an issue cluster are mentioned in many more articles.[8]

The strong correlation between number of events in an issue cluster and articles about the cluster is partly because multiple articles about an issue cluster were more likely to distinguish among the events rather than referring to them vaguely in the aggregate. But also, episodes of sustained contention generated more events and drew more articles about them. Issue clusters with hundreds of thousands of participants that got relatively little news coverage were generally centered on large peaceful demonstrations organized by broad coalitions around liberal policy issues or petition campaigns. Issue clusters with extensive coverage but fewer participants tended to be ongoing conflict episodes. Using a similar data structure to ours, Hellmeier, Weidmann, and Rød (2018) find that "salient events" (defined as events that had at least five reports) increased the number of reports of subsequent events, especially those in the same city. Their methodology would be improved if they could determine whether their finding is due to media attention to issue clusters.

Discussion and Conclusions

The same relational data structures and workflows that promote the accuracy and verifiability of protest event data also permit the construction of data sets that better represent theoretical and empirical understandings of how protests and news stories about protests are produced. Relational data structures can capture the structuring of protest events into campaigns and episodes and illuminate the processes affecting how news media cover protests. There are always source effects: identified protest events are always the product of nonrandom selection processes specific to the media sources, whether specific newspapers, Google aggregations, or social media.

Attending to the specificity of the media sources deepens the analysis. For example, in an empirical report from our data (Oliver, Lim, Matthews, and Hanna 2022), we emphasize that our mainstream newswire sources embody the "White gaze" in what they deem important to cover about the Black movement. We have ongoing data collection to compare the newswires to Black newspapers in how they portray the Black movement. Other studies, such as the comparison of movement and mainstream sources in their coverage of WTO protests in 1999 by Almeida and Lichbach (2003) or Weiner's (2009, 2011) comparison of White and Black newspaper coverage of Black parents' protests about New York schools, can be understood as accounts of what different media find interesting to convey to their audiences. Information about Chinese protests identified through AI processing of images and texts in Weibo (the Chinese equivalent of Twitter) by Zhang and Pan (2019) is constrained by the sparse information available in brief posts. Patterns of event coverage in Facebook or Twitter are shaped by corporate-controlled algorithms, bots, and trolls, in addition to individuals and movements trying to communicate their messages. Every protest event study draws on texts or images produced by intentional actors constrained by the limitations and foci of the platform they are using. Highlighting these source-specific details deepens the understanding of movements and how they are affected by media. There is no such thing as a neutral or unbiased media source for protest event data. Interpretations of protest event data are enriched by naming and theorizing the source.

Data Accuracy and Verification

Parsing event information from sources and especially matching up events between sources are difficult and error-prone processes for either machines or people. Both machines and humans have difficulty distinguishing between reports of past and contemporary events, disentangling reports of multiple events in the same source, and matching events between sources. Datasets should maintain rigorous links between events and the full texts that describe them. This is the bare minimum for verifiability. But for data accuracy, you must not only *permit* verification, but *do* the verification.

Failure to check and verify the data does not mean they are error free. Subjective perceptions that coding is unproblematic cannot be trusted without post hoc tests. We thought our data were in good shape after the first author had reviewed the coding of every event and matched events between articles using spreadsheets, but we found many errors when we reviewed this work with a better interface. We could correct the errors because we had maintained the links to the original texts, but we only found them because we reviewed the prior work. Training coders to recognize and classify events and assessing intercoder reliability for event characteristics are valuable, but do not address deduplication issues in matching events up between sources. We recognize that all data collection involves some error, but there is a huge difference between 1% and 20% or 70% error rates. Published assessments of machine-coded event data have found extraordinarily high error rates. Published intercoder reliability statistics for human-coded data imply error or inconsistency rates of at least twenty percent even in well-funded projects with highly trained coders, and those statistics do not measure deduplication errors.

Our (perhaps obsessive) desire to correct errors in our data has led us to recognize the value of providing more detailed information in the dataset. We have also learned the importance of human-machine interfaces that facilitate searches and present information to a human coder in an

accessible format. As we have moved to analysis, we have found more errors or inconsistencies that our data structures have allowed us to resolve and then reexport the corrected dataset.

Projects vary greatly in the resources available to them, but whether a project is well-funded with a dozen or more well-paid employees or solitary scholars coding their own data, all projects should plan for the joint realities of coder error and coding ambiguities. Our initial data collection was largely unfunded and relied upon undergraduates with limited training and experience; after receiving funding we spent significant time correcting errors made earlier. But even the most well-trained and conscientious human coders make mistakes, and machine coders make even more. In addition to maintaining links to sources to permit error checking, all projects, whether large or small, should budget part of their project time for reviewing at least a sample of coded data to estimate error rates. Matching events between articles is error prone: instructing coders to ignore events they believe are duplicates is bad practice. At a minimum, articles tagged as containing only duplicates should be reviewed, but it is better practice to retain the links to all articles mentioning a widely discussed event as the media coverage data itself is important, even as it is not necessary to conduct a detailed coding of every article about such an event. Providing coders with articles presorted by date, location, and content similarity improves the ability to detect duplicate events and identify related nonduplicate events. Identified events should be sorted by date, location, and issue to permit a further review for deduplication errors. Methodological reports should explicitly state how multiple reports of the same event were handled and describe procedures for reviewing and correcting errors.

Relational data structures are best for capturing the many-to-many relationships between events and news articles. Relational databases themselves are the key to data integrity controls. These data structures must be accompanied by research protocols and coding interface tools that permit and encourage the systematic review and correction of prior coding. The online appendix displays some of our interface tools.

Our project uses news sources to retrieve information about past protests in an understudied historical period, and our key recommendations are best suited for similar projects. We recognize that some projects are focused on generating immediate reports on current events and are willing to trade reduced detail and accuracy for a fast turnaround. Some basic principles seem relevant even for these. A small number of events will receive disproportionate mention in any pool of sources. There will always be errors in identifying and deduplicating events, and there should be procedures for at least assessing and reporting estimates of these errors. Although broad, high-level initial descriptive reports may provide a useful snapshot for early warning or immediate policy discussions despite undiagnosed errors, multivariate analysis of these datasets without attention to the likelihood of systematic errors and source effects is highly problematic. Web-scraping procedures that aggregate diverse sources increase the uncertainty about the selection processes underlying the data. Retaining all citation information permits post hoc analyses of source effects that may shed light on these issues.

Handling Inherent Ambiguity and Relations

Recording greater specificity and detail upfront greatly improve the ability to match events between sources, recognize issue clusters, and acknowledge ambiguity and edge cases. Data collection protocols should plan for protest events' inherent ambiguity and relationality. Instead of immediately classifying events in texts into preplanned categories, protocols should record more specific detail about events by marking text and writing open-ended descriptions that can include comments about ambiguity. Our experience across several projects involving different teams is that even the principal researchers may disagree about how to interpret an ambiguous event after an hour of discussion. These issues are unlikely to be adequately resolved by a coder staring at a confusing text and reviewing a thirty-page memo of coding rules in their first and only encounter with the text.

Data collection protocols should plan for multiplace, multiaction, multiactor, multidate events and provide some way for coders to report the connections between the components of

complex events, as well as their recognition that multiple events are about the same issue. As we have described in this article, relational data structures can capture these relations. Our project recognized these relations early but initially had only a primitive way of recording them, so we have had to do most of this work post hoc.[9] Similarly, providing event identification coders with a list of already-recognized events and issue clusters should speed up event adjudication coding.

Deep theoretical issues are present in the mundane procedures of search strings, source selection, and reading and coding texts. Attending seriously to the joint problems of verifiability and deduplication has led us to recognize the inherent ambiguity and relationality of protest as a concept. These ambiguities can be resolved only by understanding the relation of one event to other events. What exactly was the issue? Does the word "protest" refer only to verbal complaints, or was there some sort of physical action? Into how many distinct events should a coder parse a series of actions by a shifting pool of participants? How should we count the hundreds of reporters, police, and bystanders watching a confrontation between a handful of KKK and New Black Panthers? How do we count consumer boycotts, online mobilizations, petitions, lawsuits, concerted letter-writing and telephone-call campaigns? In moving toward a quantitative analysis, a researcher must address what counting is meaningful in a specific research analysis context.

THE PAYOFF: NEW DATA FOR NEW THEORIZING

The same research protocols that make event data verifiable put protest data on a sounder theoretical grounding and open the door to new research possibilities. When we truly accept that the data are about protest events *as they appeared in these sources,* we can ask new questions and offer new evidence about old questions. Attending to duplicate event reports focuses attention on the texts and the social and organizational processes undergirding the production of those texts but also allows us to distinguish between the actions of claimants and the actions of news media. This differentiates our approach from the "claims-making" approach (Koopmans and Statham 1999a, Koopmans and Statham 1999b), which appears to make no distinction between the same claims-making event reported ten times and ten distinct claims-making events.

For example, our preliminary results suggest that the volume of media coverage about an issue is affected more by the ability to mount multiple protests over time than by the number of participants. But the ability to sustain protest is endogenous, as media coverage can aid the diffusion of protest and recruitment of protesters in the positive feedback system described by Seguin (2016). We can study the role of frequently covered events in these processes.

Linking events to texts led us to attend to how events are discussed in the context of other events. This discursive structuring of events into specific-issue clusters is closely tied to a long-standing recognition that events are generally structured as campaigns and episodes, and that news reporters are trained to follow ongoing "stories" in media attention cycles. Our preliminary results suggest that capturing issue clusters provides high explanatory power in understanding patterns of media attention. Furthermore, recognizing issue clusters puts event data into dialog with qualitative historical work. It has led us, for example, to look more closely at the major episodes of protests about police violence in the 1990s and 2000s and examine how descriptions of them may have been affected by the political contexts of the cities involved. Many smaller protest event projects are parts of projects that also include qualitative historical materials. Paying attention to specific-issue clusters and sources will allow an even deeper integration of quantitative and qualitative materials.

Our recommendations about how to do the initial data collection are deeply substantive and theoretical. They flow from both a recognition of the inherent relationality of protest and the way everything we know about protest is filtered through texts that describe protests. Embracing and not discounting the difficulty of the data-collection task and the specificity of our media sources will deepen theory and allow us to generate richer data that will tell us more about how protests work.

NOTES

[1] The Mobilization in Autocracies Project treats the event report (an article mentioning an event) as the unit of analysis, the Crowd Counting Consortium project provides up to thirty URL citations for each event, and the data files supporting the Count Love project also provide all the URLs for each identified event. None of these projects explicitly describes relational data structures although the Autocracies Project uses a hierarchical approach.

[2] Protest relevant: boycott* OR press conference OR news conference OR protest* OR strik* OR rally OR ralli* OR riot* OR sit-in OR occupation OR mobiliz* OR blockage OR demonstrat* OR marchi* OR marche* NOT protestant*. Black relevant: (Black AND NOT Blacks) OR African* OR Afro*

[3] The data cleaning and deduplication processes for the Dynamics of Collective Action dataset are described in a document available online (Ku, Rafail and Wang 2009). Personal communications with PIs Susan Olzak, Sarah Soule, and John McCarthy, and then-graduate student Patrick Rafail, have established that dealing with duplicate events in DCA was a difficult problem. Initial coders were instructed to ignore duplicate events, but many were recorded, probably because there were many coders on the project who would not know what others had coded. Deduplication involved trying to match events up by date, location, and categorized issue. PDF copies of cited news sources were consulted.

[4] Twenty-two events were in a cluster that was nested within a larger cluster; twenty-five events were in two overlapping clusters and two were in three overlapping clusters.

[5] There are 216 issue clusters with only one event; four involve events in more than one cluster.

[6] In the cases of large coalitional rallies, it is possible that there were other newswire stories we did not select because of the screening for Black-relevant keywords, but overall they seemed to be treated as less newsworthy than conflictual events.

[7] The correlation between the log of the participants per month and the number of articles per month is somewhat higher, $r = .54$, $r^2 = .29$, but still substantially lower than the correlation between events and articles. Various specifications of the regression of number of articles on number of events and participants all give the same result: that number of the participants has a much weaker relation to the number of articles than the number of events.

[8] Event size has the larger effect if the regression is in the original values and cluster size (number of events in cluster) has the larger effect if the regression is in logged values of all variables. Results not shown. The two predictors are uncorrelated. We also tested interaction models that show that event size matters for physical gatherings more than for nongathering events, consistent with event size being less meaningful for nongatherings.

[9] The well-funded SPEED project reports having a coding interface that presents coders with different choices depending on their responses to initial screening questions.

REFERENCES

ACLED. 2020. “Faqs-Sourcing-Methodology. V1 February 2020.” (https://acleddata.com/acleddatanew/wp-content/uploads/2021/11/ACLED_FAQs-Sourcing-Methodology_v1_February-2020.pdf).

———. 2022, “The Armed Conflict Location and Event Data Project.” Retrieved on June 16, 2022 (https://acleddata.com/#/dashboard).

Almeida, Paul D., and Mark Irving Lichbach. 2003. “To the Internet, from the Internet: Comparative Media Coverage of Transnational Protests.” *Mobilization: An International Quarterly* 8(3): 249-72.

Althaus, Scott L., Joseph Bajjalieh, John F. Carter, Buddy Peyton and Dan A. Shalmon. 2017. “Cline Center Historical Phoenix Event Data Variable Descriptions (V.1.0.0).” Retrieved: June 30, 2017 (https://uofi.app.box.com/s/bmh9i39m6bf0vhnuebtf3ak3j6uxy2le).

Amenta, Edwin, Neal Caren, Sheera Joy Olasky and James E. Stobaugh. 2009. “All the Movements Fit to Print: Who, What, When, Where, and Why Smo Families Appeared in the New York Times in the Twentieth Century.” *American Sociological Review* 74(4): 636-56.

Amenta, Edwin, Thomas Alan Elliott and Amber C Tierney. 2016. “Political Reform and Newspaper Coverage of Us Movements in Depression, Recession, and Historical Perspective.” *Mobilization: An International Quarterly* 21(4): 393-412.

Andrews, Kenneth T., and Michael Biggs. 2006. “The Dynamics of Protest Diffusion: Movement Organizations, Social Networks, and News Media in the 1960 Sit-Ins.” *American Sociological Review* 71(5): 752-77.

Andrews, Kenneth T., and Neal Caren. 2010. “Making the News: Movement Organizations, Media Attention, and the Public Agenda.” *American Sociological Review* 75(6): 841-66. doi: 10.1177/0003122410386689.

Beissinger, Mark R. 2002. *Nationalist Mobilization and the Collapse of the Soviet State*. New York: Cambridge University Press.

Biggs, Michael. 2016. “Size Matters: Quantifying Protest by Counting Participants.” *Sociological Methods and Research* 47(3): 351-83. doi: 10.1177/0049124116629166.

Bond, Doug, J. Craig Jenkins, Charles L. Taylor and Kurt Schock. 1997. “Mapping Mass Political Conflict and Civil Society:Issues and Prospects for the Automated Development of Event Data.” *Journal of Conflict Resolution* 41(4): 553-79. doi: 10.1177/0022002797041004004.

Boschee, Elizabeth, Premkumar Natarajan and Ralph Weischedel. 2013. "Automatic Extraction of Events from Open Source Text for Predictive Forecasting." Pp. 51-67 in *Handbook of Computational Approaches to Counterterrorism*, edited by V. S. Subrahmanian. New York, NY: Springer.

Campbell, Shannon, Phil Chidester, Jason Royer and Jamel Bell. 2004. "Remote Control: How Mass Media Delegitimize Rioting as Social Protest." *Race, Gender and Class* 11(1): 158-76.

Cancian, Francesca M., and Bonnie L. Ross. 1981. "Mass Media and the Women's Movement: 1900-1977." *Journal of Applied Behavioral Science* 17(1): 9-26.

Caren, Neal. 2014. "It Is Time to Get Rid of the E in Gdelt." *BadHessian.* Retrieved 2021 (https://badhessian.org/ 2014/05/it-is-time-to-get-rid-of-the-e-in-gdelt/).

CEDS. 2013, "The Computational Event Data System". Retrieved 7/16/22, 2022 (https://eventdata.parusanalytics.com).

Chalabi, Mona. 2014. "Mapping Kidnappings in Nigeria." *FiveThirtyEight.* Retrieved 2021 (https://fivethirtyeight.com/features/mapping-kidnappings-in-nigeria/).

Chang, Paul Y., and Kangsan Lee. 2021. "The Structure of Protest Cycles: Inspiration and Bridging in South Korea's Democracy Movement." *Social Forces* 100(2): 879-904. doi: 10.1093/sf/soaa130.

Croicu, Mihai, and Nils B Weidmann. 2015. "Improving the Selection of News Reports for Event Coding Using Ensemble Classification." *Research and Politics* 2(4): 2053168015615596. doi: 10.1177/2053168015615596.

Davenport, Christian. 2010. *Media Bias, Perspective, and State Repression : The Black Panther Party*. New York: Cambridge University Press.

della Porta, Donatella, and Massimiliano Andretta. 2002. "Changing Forms of Environmentalism in Italy: The Protest Campaign on the High Speed Railway System." *Mobilization* 7(1): 59-77.

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.

Fisher, Dana R., Kenneth T. Andrews, Neal Caren, Erica Chenoweth, Michael T. Heaney, Tommy Leung, L. Nathan Perkins and Jeremy Pressman. 2019. "The Science of Contemporary Street Protest: New Efforts in the United States." *Science Advances* 5(10 eaaw5461): 1-15. doi: 10.1126/sciadv.aaw5461.

Fletcher, Bill, Jr. and Jamala Rogers. 2014. "No One Said That It Would Be Easy." *Black Scholar* 44(1): 86-112. doi: 10.5816/blackscholar.44.1.0086.

Franzosi, Roberto. 1999. "The Return of the Actor. Interaction Networks among Social Actors During Periods of High Mobilization (Italy, 1919-1922)." *Mobilization* 4(2): 131-49.

Freese, Jeremy, and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43(1): 147-65. doi: 10.1146/annurev-soc-060116-053450.

GDELT. 2021, "The Gdelt Project." Retrieved 7/16/22, 2022 (https://www.gdeltproject.org).

Gitlin, Todd. 1980. *The Whole World Is Watching: Mass Media in the Making and Unmaking of the New Left*. Berkeley: University of California Press.

Giugni, Marco, Ruud Koopmans, Florence Passy and Paul Statham. 2005. "Institutional and Discursive Opportunities for Extreme-Right Mobilization in Five Countries." *Mobilization: An International Journal* 10(1): 145-62.

Hanna, Alex. 2016. "Automated Coding of Protest Event Data: Development and Applications." Ph.D., The University of Wisconsin - Madison, Ann Arbor, MI. Retrieved from Dissertations and Theses at CIC Institutions; Dissertations and Theses, University of Wisconsin at Madison; ProQuest Dissertations and Theses Global; Sociological Abstracts; Technology Collection, 10190827.

———. 2017. "Mpeds: Automating the Generation of Protest Event Data." Deposited at SocArXiv https://osf.io/preprints/socarxiv/xuqmv.

———. 2022. "Mpeds Annotation Interface." https://github.com/MPEDS/mpeds-coder.

Hayes, Matthew, and Peter Nardulli. 2011. "The Quality and Reliability of Data Generated by Speed's Societal Stability Protocol: Mechanisms and Tests." Available at https://uofi.app.box.com/s/gtqkapslqjkyzwme96v4edwsvkfr14bi.

Hellmeier, Sebastian, Nils B. Weidmann and Espen Geelmuyden Rød. 2018. "In the Spotlight: Analyzing Sequential Attention Effects in Protest Reporting." *Political Communication* 35(4): 587-611. doi: 10.1080/10584609.2018.1452811.

Hellmeier, Sebastian, Espen Geelmuyden Rød and Nils B. Weidmann. 2019. "Coding Instructions for the Mass Mobilization in Autocracies Database, Version 2.0." (http://cybis.univ-grenoble-alpes.fr:8082/dataset/08a3a90e-0b6a-435e-b566-135c8f3b9f34/resource/c542f8d9-2499-4e5a-8bba-8f68ca53a033/download/mmad_codebook_v2-20190821.pdf).

Herkenrath, Mark, and Alex Knoll. 2011. "Protest Events in International Press Coverage: An Empirical Critique of Cross-National Conflict Databases." *International Journal of Comparative Sociology* 52(3): 163-80. doi: 10.1177/0020715211405417.

Hutter, Swen. 2014. “Protest Event Analysis and Its Offspring.” Pp. 335-67 in *Methodological Practices in Social Movement Research*, edited by Donatella della Porta. Oxford: Oxford University Press.

ICEWS. 2022, “Integrated Crisis Early Warning System (Icews).” Retrieved on 7/16/22, 2022 (https://www.lockheedmartin.com/en-us/capabilities/research-labs/advanced-technology-labs/icews.html).

Jenkins, J. Craig, and Thomas V. Maher. 2016. “What Should We Do About Source Selection in Event Data? Challenges, Progress, and Possible Solutions.” *International Journal of Sociology* 46(1): 42-57. doi: 10.1080/00207659.2016.1130419.

Koopmans, Ruud, and Paul Statham. 1999a. “Challenging the Liberal Nation-State? Postnationalism, Multiculturalism, and the Collective Claims Making of Migrants and Ethnic Minorities in Britain and Germany.” *American Journal of Sociology* 105(3): 652-96. doi: 10.1086/210357.

———. 1999b. “Political Claims Analysis: Integrating Protest Event and Political Discourse Approaches.” *Mobilization* 4(2): 203-21.

Koopmans, Ruud and Dieter Rucht. 2002. “Protest Event Analysis.” Pp. 213-59 in *Methods of Social Movement Research*, edited by Bert Klandermans and Suzanne Staggenborg. Minneapolis: University of Minnesota Press.

Koopmans, Ruud, and Susan Olzak. 2004. “Discursive Opportunities and the Evolution of Right-Wing Violence in Germany.” *American Journal of Sociology* 110(1): 198-230.

Kriesi, Hanspeter, Swen Hutter and Abel Bojar. 2019. “Contentious Episode Analysis.” *Mobilization* 24(3): 251-73.

Ku, Candy, Pat Rafail, and Dan Wang. 2009. “Dynamics of Collective Action Data Cleaning Procedures.” http://www.stanford.edu/group/collectiveaction/CLEANING%20DOCUMENT%20Nov.%209,%202009.docx

Leung, Tommy, and L. Nathan Perkins. 2021. “Counting Protests in News Articles: A Dataset and Semi-Automated Data Collection Pipeline.” https://arxiv.org/abs/2102.00917.

Lorenzini, Jasmine, Hanspeter Kriesi, Peter Makarov and Bruno Wüest. 2022. “Protest Event Analysis: Developing a Semiautomated Nlp Approach.” *American Behavioral Scientist* 66(5): 555-77. doi: 10.1177/00027642211021650.

Makarov, Peter, Jasmine Lorenzini, and Hanspeter Kriesi. 2016. “Constructing an Annotated Corpus for Protest Event Mining.” Pp. 102-07 in *Proceedings of the First Workshop on NLP and Computational Social Science*, edited by D. Bamman, A. S. Doğruöz, J. Eisenstein, D. Hovy, D. Jurgens, B. O’Connor, A. Oh, O. Tsur and S. Volkova. Austin, TX: Association for Computational Linguistics https://aclanthology.org/W16-5613/.

Maney, Gregory M., and Pamela E. Oliver. 2001. “Finding Collective Events: Sources, Searches, Timing.” *Sociological Methods & Research* 30(2): 131-69. doi: https://doi.org/10.1177%2F0049124101030002001.

Martin, Andrew W., Patrick Rafail and John D. McCarthy. 2017. “What a Story?”. *Social Forces* 96(2): 779-802. doi: 10.1093/sf/sox057.

McAdam, Doug, Sidney Tarrow and Charles Tilly. 2001. *Dynamics of Contention*. New York: Cambridge University Press.

Mediacloud. 2020. “Cliff: Extract Named Entities and Geoparse the News.” https://github.com/mediacloud/cliff-annotator.

Morris, Aldon. 1984. *The Origins of the Civil Rights Movement: Black Communities Organizing for Change*. New York: Free Press.

Myers, Daniel J. 2000. “The Diffusion of Collective Violence: Infectiousness, Susceptibility, and Mass Media Networks.” *American Journal of Sociology* 106(1): 173-208. doi: https://doi.org/10.1086/303110.

Napoles, Courtney, Matthew R. Gormley and Benjamin Van Durme. 2012. “Annotated English Gigaword.” edited by L. D. Consortium. https://catalog.ldc.upenn.edu/LDC2012T21.

Nardulli, Peter F., and Matthew Hayes. 2011. “Transforming Textual Information on Events into Event Data within Speed.” Urbana-Champaign, IL: Cline Center for Democracy. https://uofi.app.box.com/s/y8nhzmzfim774flnvzro26rooz3l5tdx.

Nardulli, Peter F., Scott L. Althaus and Matthew Hayes. 2015. “A Progressive Supervised-Learning Approach to Generating Rich Civil Strife Data.” *Sociological Methodology* 45(1): 148-83. doi: 10.1177/0081175015581378.

Oliver, Pamela, Chaeyoon Lim, Morgan Matthews and Alex Hanna. 2022. “Black Protests in the United States, 1994-2010.” *Sociological Science* 9(May): 275-312. doi: DOI: 10.15195/v9.a12.

Olzak, Susan. 1989a. “Labor Unrest, Immigration, and Ethnic Conflict in Urban America, 1880-1914.” *American Journal of Sociology* 94(6): 1303-33.

———. 1989b. “Analysis of Events in the Study of Collective Action.” *Annual Review of Sociology* 15: 119-41.

Ortiz, David G., Daniel J. Myers, N. Eugene Walls and Maria-Elena D. Diaz. 2005. "Where Do We Stand with Newspaper Data?" *Mobilization* 10(3): 397-419.

Rafail, Patrick. 2018. "Protest in the City: Urban Spatial Restructuring and Dissent in New York, 1960–2006." *Urban Studies* 55(1): 244-60.

———. 2019. "Who Is Doing What and Where? Patterns of Contention in 20 U.S. Cities, 1996-2006." Paper presented at the Social Science History Association, November 2019, Chicago.

Rafail, Patrick, John D. McCarthy and Samuel Sullivan. 2019. "Local Receptivity Climates and the Dynamics of Media Attention to Protest." *Mobilization: An International Quarterly* 24(1): 1-18. doi: 10.17813/1086-671x-24-1-1.

Ratliff, Thomas N. 2011. "On the Stage of Change: A Dramaturgical Approach to Violence, Social Protests, and Policing Styles in the U.S." PhD, Sociology, Virginia Polytechnic Institute and State University, https://vtechworks.lib.vt.edu/handle/10919/28449.

———. 2013. "A Dramaturgical Approach to Protest Policing in the United States: Actors, Enemies, the Stage, and Performance." Paper presented at *American Sociological Association* Annual Meeting, Philadelphia.

Rucht, Dieter and Friedhelm Neidhardt. 1998. "Methodological Issues in Collecting Protest Event Data: Units of Analysis, Sources and Sampling, Coding Problems." Pp. 65-89 in *Acts of Dissent: New Developments in the Study of Protest*, edited by D. Rucht, R. Koopmans and F. Neidhardt. Berlin: Edition Sigmas Rainer Bohn Verlag.

Schrodt, Philip A., and Deborah J. Gerner. 1994. "Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982-92." *American Journal of Political Science* 38(3): 825-54. doi: 10.2307/2111609.

Schrodt, Philip A. 2012. "Precedents, Progress, and Prospects in Political Event Data." *International Interactions* 38(4): 546-69. doi: 10.1080/03050629.2012.697430.

Schrodt, Philip A., and David Van Brackle. 2013. "Automated Coding of Political Event Data." Pp. 23-49 in *Handbook of Computational Approaches to Counterterrorism*, edited by V. S. Subrahmanian. New York, NY: Springer.

Seguin, Charles. 2016. "Cascades of Coverage: Dynamics of Media Attention to Social Movement Organizations." *Social Forces* 94(3): 997-1020. doi: 10.1093/sf/sov085.

Snyder, David, and Charles Tilly. 1972. "Hardship and Collective Violence in France, 1830 to 1960." *American Sociological Review* 37(5): 520-32.

Soule, Sarah, Doug McAdam, D. McCarthy John and Susan Olzak. 2009. "Dynamics of Collective Action." http://web.stanford.edu/group/collectiveaction/cgi-bin/drupal/.

Strawn, Kelley D. 2008. "Validity and Media-Derived Protest Event Data: Examining Relative Coverage Tendencies 114 Mexican News Media." *Mobilization* 13(2): 147-64.

Tarrow, Sidney. 1988. *Democracy and Disorder: Politics and Protests in Italy, 1965-1975*. New York: Oxford University Press.

———. 1989. *Struggle, Politics and Reform: Collective Action, Social Movements, and Cycles of Protest*. Ithaca, NY: Center for Institutional Studies, Cornell.

Taylor, Charles Lewis. 2013. "Data Quality for Measuring Political Protest and Government Change." *All Azimuth: A Journal of Foreign Policy and Peace* 2(2): 23-29.

Taylor, Keeanga-Yamahtta. 2016. *From #Blacklivesmatter to Black Liberation*. Chicago: Haymarket Books.

Tilly, Charles, Louise Tilly and Richard Tilly. 1975. *The Rebellious Century, 1830-1930*. Cambridge: Harvard University Press.

Van de Rijt, Arnout, Eran Shor, Charles Ward, and Steven Skiena. 2013. "Only 15 Minutes? The Social Stratification of Fame in Printed Media." *American Sociological Review* 78(2): 266-89. doi: 10.1177/0003122413480362.

Vliegenthart, Rens, and Stefaan Walgrave. 2012. "The Interdependency of Mass Media and Social Movements." in *The SAGE Handbook of Political Communication*, edited by H. A. Semetko and M. Scammell. London: SAGE Publications Ltd.

Wada, Takeshi. 2004. "Event Analysis of Claim Making in Mexico: How Are Social Protests Transformed into Political Protests?" *Mobilization: An International Journal* 9(3): 241-57.

Walgrave, Stefaan and Rens Vliegenthart. 2010. "Why Are Policy Agendas Punctuated? Friction and Cascading in Parliament and Mass Media in Belgium." *Journal of European Public Policy* 17(8): 1147-70. doi: 10.1080/13501763.2010.513562.

Wang, Wei, Ryan Kennedy, David Lazer and Naren Ramakrishnan. 2016. "Growing Pains for Global Monitoring of Societal Events." *Science* 353(6307): 1502-03.

Weidmann, Nils B., and Espen Geelmuyden Rød. 2019. "Coding Protest Events in Autocracies." Pp. 35-60 in *The Internet and Political Protest in Autocracies*, edited by N. B. Weidmann and E. G. Rød. Oxford: Oxford University Press.

Weiner, Melissa F. 2009. "Elite Versus Grassroots: Disjunctures between Parents' and Civil Rights Organizations' Demands for New York City's Public Schools." *Sociological Quarterly* 50(1): 89-119. doi: 10.1111/j.1533-8525.2008.01134.x.

———. 2011. "All the News That's Fit to Print? Silence and Voice in Mainstream and Ethnic Press Accounts of African American Protest." *Research in Social Movements, Conflicts & Change* 31: 297-327. doi: 10.1108/S0163-786X(2011)0000031012.

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.