International Journal of Sociology, 46: 42–57, 2016

ISSN: 0020-7659 print/1557-9336 online
DOI: 10.1080/00207659.2016.1130419

Routledge
Taylor & Francis Group

What Should We Do about Source Selection in Event Data? Challenges, Progress, and Possible Solutions

J. Craig Jenkins

Department of Sociology, Ohio State University

Thomas V. Maher

Department of Sociology, University of Arizona

The prospect of using the Internet and other Big Data methods to construct event data promises to transform the field but is stymied by the lack of a coherent strategy for addressing the problem of selection. Past studies have shown that event data have significant selection problems. In terms of conventional standards of representativeness, all event data have some unknown level of selection no matter how many sources are included. We summarize recent studies of news selection and outline a strategy for reducing the risks of possible selection bias, including techniques for generating multisource event inventories, estimating larger populations, and controlling for nonrandomness. These build on a relativistic strategy for addressing event selection and the recognition that no event data set can ever be declared completely free of selection bias.

Keywords analytical strategies; event data methods; selection bias

The development of electronic online news archives, online activist sites, and automated techniques for computer coding of large volumes of text promises to transform our ability to describe and analyze political events and contentious politics (e.g., Almeida and Lichbach 2003; Bond et al. 1997; Chojnacki et al. 2012; Earl and Kimport 2011; Hanna 2014; Jenkins et al. 2014; King and Lowe 2003; Schrodt 2012; Shellman 2008). But, like other attempts to harness Big Data in the social sciences, this promise confronts major questions about its ability to meet conventional standards of representativeness and reliability. The two most recent

J. Craig Jenkins is a professor of sociology, political science, and environmental science at Ohio State University. He has received a Fulbright Professorship at the Peace Research Institute of Oslo (PRIO), a Leiv Eiriksson Mobility Fellowship from the Norway Research Council, and the Robin M. Williams Jr. Award for Distinguished Contributions to Scholarship, Teaching and Service from the Section on Peace, War and Social Conflict of the American Sociological Association. He has written or edited 4 books and over 100 research articles addressing social movements, protest theory, event data methods, and political economy.

Thomas V. Maher is a postdoctoral researcher in sociology at the University of Arizona. His research focuses on the intersection between social movements, organizations, and political sociology. He is primarily interested in how states and organizations control populations and how these conditions influence collective action. He has published work on these issues in outlets such as *Journal of Peace Research, American Sociological Review*, and *Mobilization*.

Address correspondence to J. Craig Jenkins, Department of Sociology, Ohio State University, 238 Townshend Hall, 1885 Neil Ave. Mall, Columbus, OH 43210. E-mail: jenkins.12@osu.edu

reviews of the literature come to quite different conclusions about the severity of the problem. Reviewing over thirty years of research on the issue, Earl et al. (2004: 77) conclude that selection bias in news data (and protest event data in particular) is comparable to that found in survey research and studies using official crime data: "Researchers can effectively use such data and that newspaper data does not deviate markedly from accepted standards of quality." In contrast, Ortiz et al. (2005: 397) review much of the same literature and come to a more pessimistic conclusion: "Newspaper data often do not reach acceptable standards for event analysis and that using them can distort findings and misguide theorizing. Furthermore, media selection biases are resistant to correction procedures largely because they are unstable across media sources, time and location."

We assess this debate over the reliability of event data and news selection in particular by addressing four sets of questions. First and foremost, how serious is the problem of source selection? Do specific news or information sources, types of events, contexts, and other features of events and news operations affect the severity of news selection? How large and unstable across time and space are these selection dynamics? Second, what is the best strategy for dealing with these problems? Should we conceptualize this in absolute terms as a bias against a known population? Or should we treat this as relative inference problem? Third, does news selection affect the results of causal inference in empirical studies using event data? Fourth, are there data collection procedures and analytic methods that can help address the problem?

It is important to state at the outset that for some purposes news selection is not a major issue. If the aim of a study is to say something about how variables are related to one another or to understand the dynamics in a particular case, then a high quality purposive sample is likely fine. If, however, the aim is to say something generalizable and to draw causal inferences, then knowing the extent and nature of any selection bias and, if possible, correcting for it is important. It is also important to note that if the study uses protest as an independent variable, for example, to predict policy change (McAdam and Su 2002) or governmental instability (Ward et al. 2013a), then mass media coverage is likely a central mechanism in the resulting political process, which reduces the severity of the problem. If, however, event data is the dependent variable, then addressing the question of random selection is critical.

THE PROBLEM OF SOURCE SELECTION

At its core, the problem of news selection stems from the fact that we lack comprehensive population-level inventories of all "real world" political events from which we could draw random samples. The universe is, strictly speaking, unknown. Even when we draw on a well-designed event inventory based on multiple sources or use authoritative sources such as police records, we ultimately do not know how this inventory relates to the full population of "real world" events. The resulting bias may be small or large but the problem is that we have no direct way to randomly sample an unknown universe.

One crude starting point for evaluating the source selection question is the proportion of protests that are covered in one source compared to another or to a multisource inventory or a seemingly authoritative source like police records. While a few studies have found little news selection (e.g., Martin's [2005] comparison of strike reportage in the *New York Times* versus the *Daily Labor Report*, a specialized labor newspaper), the majority of studies have found

significant discrepancies. In the typical study, a base code, such as an integrated multisource event inventory or a single authoritative source such as police records, is used for comparison, and a single news source covers no more than 20–40 percent of the fuller inventory of events (the rate is often significantly lower). For example, Barranco and Wisler (1999) found that nearly 50 percent of protests identified in police reports were covered in a Swiss newspaper. In a similar study, Hocke (1998) found that the local paper in Freiburg, Germany, covered about 38 percent of all protests covered in police records, that only a few events in news sources were not in police records (making this by far the most complete single source), and that three national German newspapers covered only 4.6 percent of the Freiberg protests. By contrast, Fillieule (1998) found that only 2–3 percent of the police-reported protests were also reported in national French newspapers.

Some argue that state-owned media and ideological or partisan news sources have a political stake in reporting and will either ignore antiregime protests or, for partisan sources, cover only those events that are supportive of their ideological positions. In their study of French protests, Barranco and Wisler (1999) found that conservative papers were less likely to report violent protests, apparently because they wanted to prevent "copycat" violence. In a study comparing mainstream and partisan news sources in their coverage of left-wing and right-wing movement events, Rohlinger et al. (2012) found that mainstream newspapers were more likely to report on events organized by groups with professional staff (regardless of ideology), whereas partisan news sources (both conservative and liberal) tend to report events of like-minded groups without differentiating between professional and voluntary groups. Professional staff provided not only legitimacy and credibility to mainstream sources but also a liaison for collecting information.

A similar picture is provided by Davenport's (2010) study of the mainstream versus the movement press coverage of the activities of the Black Panther Party in the 1960s. In most respects, these media focused on quite different aspects of Panther Party activities. Mainstream media focused on the coercive challenges of the black dissidents to the state and to the court procedures launched against activists, making the state appear to be in control and dissidents to be violent. In contrast, the black power movement press provided more coverage of noncontentious party activities and of the police coercion that targeted activists, making the state appear repressive and antagonistic. Which is more accurate? Davenport (2010) argues that this is an invalid question and that a more accurate picture is provided by integrating these two event catalogs while respecting the perspectives of the different sources.

One study that contravenes this conclusion about types of news sources is McCarthy et al. (2008) about protest reports in state-owned versus private for-profit newspapers in Minsk, Belarus, during the immediate postcommunist transition. In this study, four individual newspapers each reported about 30 percent of the protests found in police records and, when combined, covered around 38 percent of the police-reported events. Comparing state-owned versus for-profit papers, they found no selection differences. Despite different organizational incentives for reporting, conventional news market criteria appeared to be operating in the state-owned press.

An additional constraint on reportage is the density of other competing newsworthy events in the same time period relative to the space available for reporting. Protests must compete with other protests and other kinds of news for coverage. Sometimes called the "newshole" because it depends on the amount of news space available in any particular news source, a smaller proportion of events will typically be reported when there is a great deal of newsworthy activity (Hocke 1998; Myers and Caniglia 2004; Oliver and Maney 2000; Oliver and Myers 1999). While the size of the

"newshole" might be somewhat elastic, the ultimate size of the available "print space" is a constraint, a fact that, as we discuss below, can be controlled for in regression analyses.

More subtle forms of selection may also stem from the business model of for-profit media. The standard assumption is that commercial media have an interest in reporting large, controversial, and unusual protests that are of interest to their readership (see more below). However, some argue that events that threaten the flow of profits or challenge the existing power structure are less likely to be covered (Boycoff 2006; Herman and Chomsky 1988; Parenti 1993). Beyond news selection, corporate interests and advertisers can also influence the framing of events that are covered, what topics get attention, and what perspectives are avoided or emphasized. Fitting with the ongoing thematic agenda of the newspaper is important. In their study of the news coverage of local environmental groups in 11 major daily North Carolina newspapers, Andrews and Caren (2010) found that groups that used conventional advocacy tactics and avoided a confrontational strategy were more likely to be reported. Furthermore, groups that emphasized state- and national-level economic issues, especially farming, and avoided ecology and land preservation topics, were more likely to be reported. In a study of local newspaper reports of protests in Madison, Wisconsin, Oliver and Maney (2000) found that protests targeted at legislative issues were more likely to be covered while the state legislature was in session.

In general, the ongoing concentration of media ownership should exacerbate the problem of the "newshole" by limiting the diversity of news outlets and perspectives that are available in the larger mass media arena. At the same time, the growing importance of newswires and other online sources of news and event reports should make it easier to secure coverage and to integrate multiple sources.

Media coverage is also limited by the routines, infrastructure, and resources used by media to collect information. In an early study, Danzger (1975) showed that the number of riots reported for cities was a function of the presence of a wire service office in that city. Reporters often turn to government officials for information on developing events. They also have beats and routines that put certain events and places in their paths while other events are too remote from reporters' normal activities to be covered (Oliver and Myers 1999). This also means that the chance of coverage is increased for routinized events that conform to expectations or occur at anticipated dates, times, and central locations (Oliver and Myers 1999; Oliver and Maney 2000). Movements with a professional staff and a media effort that attend to these news routines are more likely to have their events reported (Andrews and Caren 2010; Rohlinger et al. 2012). As newspapers downsize their reporting staff and rely more heavily on centralized newswire services for content, certain kinds of events, among them social movement activities, may be less likely to be covered. The only counterbalancing force is the expansion of online sources, including news digests, that make integrated news sources and automated coding more feasible.

In addition to source selection and media features, media are more likely to report particular types of events. The basic logic is one of "newsworthiness," that is, what makes an event worthy of being reported. In general, the most newsworthy are events that are unusual, that stand out in terms of size, violence, contentiousness, and other features. Specifically, the following features appear to create a higher likelihood of reportage:

1. The size of the event (i.e., the number of participants) (Barranco and Wisler 1999; Fillieule 1998; Herkenrath and Knoll 2011; McCarthy et al. 2008; McCarthy, McPhail, and Smith 1996; Mueller 1997; Oliver and Myers 1999);

2. The geographic distance between the event and the media source, especially its reporting market and audience (Barranco and Wisler 1999; Fillieule 1998; Herkenrath and Knoll 2011; Hocke 1998; McCarthy et al. 1996; Mueller 1997; Strawn 2008);
3. The extraordinariness of the event in terms of unruliness, arrests, violence, the presence of counterdemonstrators, and the flamboyance of events (Barranco and Wisler 1999; Mueller 1997; Myers and Caniglia 2004; Oliver and Myers 1999; Snyder and Kelly 1977);
4. The fit with the ideological stance and the thematic coverage priorities of the media source (Andrews and Caren 2010; Barranco and Wisler 1999; Mueller 1997; Oliver and Maney 2000; Rohlinger et al. 2012; but see McCarthy et al. 2008); and
5. The legitimacy and professionalism of the event sponsor, including the presence of celebrities and other sources of legitimacy (Andrews and Caren 2010; McCarthy et al. 2008; Oliver and Maney 2002; Rohlinger et al. 2012 Snyder and Kelly 1977).

Paralleling these general patterns for events, Amenta et al. (2009) find that *New York Times* mentions of social movement families are greater for larger, better–organized, and disruptive movements that use protest and those with an enforced governmental policy in place. They also find that having political allies in power does not influence selection.

These studies have led to the general conclusion that multisource event inventories have substantial advantages over single source inventories and that the inclusion of official or authoritative sources such as police reports greatly improves representativeness. A note of caution, however, is warranted. In their study of urban riot selection, Myers and Caniglia (2004) found that adding *Washington Post* to *New York Times* reports actually magnified some of the selection bias found in the latter when compared against a more complete inventory generated through extensive multisource data collection. Simply adding an additional source may not always enhance representativeness.

The logic of these studies is that by identifying the nature of selection bias in a particular source, one can take this information into account in drawing conclusions from a particular study using these data. In other words, when evaluating a study of a particular movement or form of activity that is selectively underreported (or overreported relative to other events or movements), one should be more cautious in accepting results. While in principle this is valid, a key question that lies behind this approach has not been addressed. Is the logic for the comparison based on an absolute standard, that is, the base code is seen as constituting the full population of events? Or is this a relative inference problem in which one is attempting to identify bias in a single source relative to other partial and limited sources, so as to reduce the likelihood of source-specific inference error? While this might seem like splitting hairs, it is critical to deciding how to proceed. If an absolute base code is possible, then we should focus on constructing it. But, if as we contend, there is no ultimate base code, then the best approach is to devise methods to assess the contributions of inevitably partial and limited data sets.

In this regard, it is well to keep in mind the conclusion of Oliver and Myers (1999: 48) about the completeness of police records, which have often been seen as authoritative sources. They found police records to be "kept unsystematically" and to vary widely in terms of their completeness and "details about the numbers, actions, identities or issues of protestors." As they conclude, "all record sources must be treated as incomplete. Different record sources must

be assessed against each other to determine their logic of inclusion and exclusion of events." In other words, all sources are partial and limited.

A parallel study that strongly recommends this relative inference approach is that by Davenport and Ball (2002). In comparing death and "disappearance" estimates stemming from state terror in Guatemala between 1977 and 1996, they argue that three seemingly independent sources—newspapers, human rights documents, and interviews conducted by a human rights organization—provide different components of the overall pattern. Newspapers tend to focus on urban environments and are most complete when the highest numbers of killings are occurring and the regime is not highly restrictive with regard to press freedoms and operations. Human rights organizations are most complete when large numbers of individuals are being killed and when political openness and press freedoms are limited. Interviews highlight rural areas as well as more recent events where memories are clearer. It is also worth pointing out that human rights organizations may have a stake in overreporting since this testifies to their worth. Davenport and Ball warn against being "dismissive of information or research that is based on one source; rather, we should endeavor to understand the limitations of all single-course analyses from a juxtaposition across distinct types" (2002: 447). They recommend disaggregating data along geographic units and time periods, where they found the greatest discrepancies, as well as qualifying conclusions based on these dimensions.

The common call to add more, and more diverse, sources to address news selection often overlooks the question of what additional sources add and whether they might amplify selection bias. Adding additional data sets needs to be evaluated in terms of what they add to the representativeness of samples. This is not simply a question of cost but also of representativeness. In addition, we need to know what the procedures are for identifying duplicate reports of the same "real world" event. Adding additional news sources increases the likelihood of multiple reports about the same event. How is a matched report identified? Does one integrate the additional information in multiple reports into the data? Finally, we need to recognize that "authoritative" nonmedia sources such as police records, the state, and nongovernmental organizations (NGOs) are likely to have their own selection biases, which must also be taken into account (Hafner-Burton and Ron 2009; Oliver and Myers 1999).

DOES SOURCE SELECTION AFFECT STUDY RESULTS?

A second more precise way of answering the question about the severity of the source selection is to examine the temporal and geographic stability of the selection process and to assess whether different sources produce different analytic results. In other words, what is the predictive validity of different sources?

Unfortunately, only a few studies have actually addressed this question. In their study, Davenport and Ball (2002) show large differences in the selection process by different types of information sources that vary significantly across time and space. The size of these discrepancies is sometimes huge with human rights sources reporting as much as 50 times the number of estimated deaths as in newspapers. Moreover, the temporal match of these deaths is quite different. The human rights organizations and the interviews put most of the killing in a one-year spike while the newspapers show greater consistency across time. Nonetheless, Davenport

and Ball (2002) conclude that each source is valuable and that the best approach is to take these differences into account and generalize to specific contexts.

A quite different conclusion is reached by Ortiz et al. (2005: 397), who conclude that media selection dynamics are "resistant to correction procedures largely because they are unstable across media sources, time and location." In a regression analysis of the coverage of urban riots in two years (1968 and 1969) in the *New York Times* versus a larger multisource inventory, Ortiz et al. (2005: 410) find that only two predictors out of five are temporally consistent ("in NY State" and "College or University") and that five predictors ("Distance from NYC," "Event Intensity * Distance," "Proportion Black," "Event Density," and "Day of Week") are inconsistent, showing statistical significance in only one or the other year. The most critical is "Proportion Black," which has been a controversial finding in earlier studies. If "Proportion Black" is relevant only in certain years, then its effects may be time-dependent and, in turn, may not be replicated in other samples. However, it is also worth noting that all variables in question maintained the same signs and "Proportion Black" was close to conventional significance levels ($p = .11$), so the risk of inference error might actually be relatively limited.

A third study addresses the question of temporal stability. In their study of postcommunist protests in Minsk, Belarus, McCarthy et al. (2008) find that the standardized coefficients of event size and sponsorship for predicting inclusion relative to police records are identical for the four newspapers. In other words, the selection process was identical across three politically different time periods (i.e., the postperestroika crisis to the fall of the USSR; the three-year parliamentary republic; and the one-and-a-half-year presidential republic). They conclude that news selection displays "remarkable stability through the volatile transition and across four very diverse newspapers" (McCarthy et al. 2008: 142). Furthermore, they note that their findings about protest size are consistent in other country studies (e.g., Switzerland, the United States, etc.), suggesting a patterned selection process.

A more convincing answer comes from studies that compare prediction results using different independently collected samples. In a study of political violence in Northern Ireland, White (1993) compares political violence deaths reported in the *New York Times Index* (*NYT Index*) with those generated by the *Agenda* database published by the Irish Information Partnership (a local NGO documentation project) for August 1969–December 1980. Although *Agenda* produces a higher total fatality count (2,062 fatalities vs. 1,448 in *NYT Index*), the regression results using four independent variables (a lagged endogenous term, regime repressiveness, a truce period dummy, and percent unemployed) were virtually identical. Only the truce dummy differed, showing significance in the *Agenda* analysis but not in the *NYT Index* analysis. White concludes that these "are basically identical to the statistical inferences produced by a comparable measure from the *Agenda* database" (1993: 583), but notes that this is a well-covered conflict with significant U.S. news interest. "If Northern Ireland were a Third World country, reliable coverage might not obtain" (ibid.).

A different answer is given by Myers and Caniglia's (2004) analysis of urban riots as reported in the *New York Times* (*NYT*) and the *Washington Post* versus those provided by a multisource inventory constructed from hundreds of local newspapers. The *NYT* reported 37.5 percent of all events and the *NYT–Washington Post* 44.7 percent of 1,114 riots. In line with the above-discussed selection mechanisms, *NYT* and *NYT–Washington Post* coverage were enhanced by proximity to New York City, event intensity (number of deaths), occurring in a college (vs. a secondary school), black population size, and negatively by event density.

More significantly, they also compared Cox regressions of the risk of a riot based on city characteristics, comparing the results from the combined data set versus *NYT* and *NYT–Washington Post* coverage. While most of the effects are statistically significant in all three equations, one variable is unique to the combined data set (black unemployment, which is on the margin of being statistically significant in the *NYT–Washington Post* data set [$p = .102$]) and another (black median income) is statistically significant only in the combined and *NYT* data sets. Furthermore, they argue that the sizes of the coefficients relative to their standard errors are much larger in the *NYT* data set, suggesting that this "can mean substantial differences in the interpretation of the results, and in other circumstances (depending on the size of the original coefficients), could produce completely different findings for a variable" (Myers and Caniglia 2004: 534).

While these are differences, it is well to keep in mind that all except one variable showed statistical significance. The variable in question—black unemployment—is theoretically important because it seems to tap relative deprivation as a factor, which other studies have not detected, but this is only one of six statistically significant factors. Overall, their main point is clear: multisource inventories provide a stronger basis for inference, if only because of their larger and presumably more complete coverage.

CAN WE FIX THIS WITH THE INTERNET?

The development of the Internet and the availability of integrated online news archives coupled with the development of computational tools for coding large amounts of electronic text into event data promises the possibility of constructing multisource event data sets (Bond et al. 1997; Gerner et al. 1994; Hanna 2014; Jenkins et al. 2014; King and Lowe 2003; Leetaru and Schrodt 2013; Shellman 2008; Schrodt 2012). Some projects, such as World-wide Integrated Crisis Warning System (ICEWS; http://www.lockheedmartin.com/us/products/W-ICEWS/W-ICEWS_overview.html) and the Phoenix Data Project (http://phoenixdata.org) have developed highly sophisticated programs for integrating hundreds of diverse sources to construct international event data and are beginning to release their data for secondary analysis. While these are major developments, major challenges still exist.

First, no one knows what the selection pattern is in such multisource data sets. Despite using hundreds of sources, it is still possible that there is significant selection in these event inventories. In addition to looking at specific source bias, along the lines discussed above, we should also make use of "capture/recapture" or multiple systems estimation (MSE) methods to project the possible larger universe of political events. Basically these methods realize that we will never have more than partial and potentially selective samples and, by using sophisticated projection methods applied to multiple event inventories, estimate what might be a larger universe of such events (Seybolt, Aronson, and Fishhoff 2013).

Second, there is possible temporal and spatial instability of online news archives that are difficult to assess (e.g., Ortiz et al. 2005). News integrators periodically shift the archives available online to fit archive space, perceived newsworthiness, and so on, which makes it important to know more about temporal and spatial bias.

Third is the problem of coding reliability. Automated coding methods have made great improvements but coding accuracy remains an issue. In a comparison of machine versus human

coding, King and Lowe (2003) found that machine coding using the VRA Knowledge Manager parser was comparable in accuracy to human coding—in some cases as low as 25–50 percent for the detailed event types (e.g., political graffiti) and up to 55–70 percent for the more generic cue category events (e.g., protest demonstrations). In general, the simpler the coding scheme, the greater the accuracy. But many event data analysts will want better than 60 percent coding accuracy before they will use such data, so further work may be necessary before we have the level of accuracy desired.

Fourth is the problem of resolving multiple reports of the same event. In electronic news sources, duplicate reports may be due to reprints, which are common in newswires, or multiple reports of the same event as additional details become available or as corrections to earlier stories are issued. News archives also contain news digests that repeat summaries of previously distributed stories. And, of course, the larger the number of news sources, the more likely that there will be independent reports of the same event from multiple news sources. When trying to measure trends in behavior over a baseline, these duplicates represent a major challenge that grows with the size and complexity of news archives.

To give some idea of how serious the problem could be, there is an online discussion of GDELT, an automated multisource event database that makes use of Googlenews.com to create "near real-time" daily updates of violent and other events along with additional conflict indicators (Leetaru and Schrodt 2013). Several analysts using Global Database of Events, Language and Tone (GDELT) for humanitarian early warning have noted that GDELT does not currently clean or mark for duplicate reports. As a result, a naive user might take literally the 649 kidnappings reported for Nigeria during the month after April 14, 2014. Actually this is the number of news reports about the same mass kidnapping by the Boko Haram that GDELT located (causalloop.blogspot.com/2014/05/how-bad-are-duplication-problems-in.html). Other systems (e.g., ICEWS [Ward et al. 2013b]) have developed strong methods for identifying duplicate reports. Analysts need to establish clear minimum criteria for defining an event match and decide how to deal with integrating information. One possibility is to provide details on source-specific information, leaving the final decision up to the analyst (for an example, see Chojnacki et al. 2012).

The Internet also creates the possibility of coding activist Web sites, blogs, and the like, which may provide event summaries. In a unique study, Almeida and Lichbach (2003) compared the protest counts on activist Web sites with those from local, national, and international newspapers and news wires. Using as their focus the December 1999 protests against the World Trade Organization summit in Seattle, they find that activist Web sites are more complete than any other source, reporting almost half of a larger multisource inventory, and are less selective regarding event intensity, reporting more protests that are smaller and nonviolent. Further, they are more likely to report protests at the local, national, and international levels. However, for local social movement organization (SMO) protests located in Seattle, the international news archive LexisNexis reported the greatest number of events and, for national events outside of Seattle, the *New York Times* reported the smallest number.

However, not all activist Web sites are equally valuable. To build this database, they consulted 20 Web sites. Web sites with a special news section were the most valuable, providing event chronologies, archiving messages and eyewitness reports, and maintaining electronic hyperlinks to news articles elsewhere. It was a significant challenge to develop procedures to confirm that reported events actually occurred, for example, by comparing Web reports against

local media coverage. Of course, not all protest campaigns will have Web sites that are constantly updated and maintained.

A final issue in using the Internet involves how to assess the representativeness of online sources. Almeida and Lichbach (2003) seem to have used a "network" approach to identify their population by starting with a small number of "seed sites" and using hyperlinks to find additional activist Web sites. While this is effective when dealing with a compact and strongly networked protest campaign, it may not work in more diffuse movements or where hyperlinks are less meaningful (Ackland 2009). In a comparison of methods for creating population estimates for studying Internet political activism, Earl (2013) finds that a "reachable" Web sites approach is the most effective. The objective is not to identify all online content relevant to protest, which is impossible given the complexity of the Internet, but to identify all content that could be located by a user who did not already know its location. By making assumptions about how users locate Web sites (largely through searches [e.g., Google] or navigating links from sites they have already found), one develops a list of pretested search terms for the topic of interest and then deploys multiple search terms (6–14 depending on the topic) to identify online instances of protest tactics. Each term generates 1,000+ results, which concatenated generates 6,000 to 14,000 results. Once cleaned for duplicates, this list of Web sites then constitutes the sampling frame of public Web sites from which a random sample can then be drawn for detailed data collection and analysis. This performed better than either random sampling a large list of movement organizations (e.g., provided by the *Encyclopedia of Associations*) or using expert knowledge of a social movement family, by identifying 33–40 percent more Web sites for examining offline protest reports. The organizational sampling approach overrepresents older established SMOs that get listed in the *Encyclopedia.* The expert knowledge approach has no information on the larger population. The "reachable Web sites" approach is, of course, limited by the search technology and the timing of the search but it provides a more systematic way of thinking about how to sample the Internet.

ARE THERE ANALYTIC FIXES?

A final question is whether source selection can be treated through analytic methods. Scholars have developed a number of approaches for modeling nonrandom error, which, in principle, can be applied to event data. One approach is to treat the error as endogenous. Newspaper data are often viewed as produced by a mass media system that is a central part of the interactions between state, various publics, and policy outcomes (Koopmans 2004; Oliver and Maney 2000). In other words, the selection is likely influenced by (i.e., endogenous to) this process. But the media also constitute a separate actor, and so it is important to clearly delineate the roles of the state, movements, and media in the process.

The most common approach to modeling endogenous nonrandom sampling error is Heckman models (Heckman 1979; Betz 2013). Heckman models are two-stage models where the researcher treats "unobserved selection factors as a problem of specification error or a problem of omitted variables, and correct(s) for bias in the estimation of the outcome equation by explicitly using information gained from the modeling of sample selection" (Guo and Fraser 2014: 86). This approach treats the omission (intentional or otherwise) of small, spontaneous,

or unrecognized events as a truncation problem that is endogenous to the model because the factors that influence coverage—as noted above—also influence event occurrence. If, using this approach, sampling selectivity is detected (i.e., the rho is not zero), treatment effect models are appropriate.

There are two notable problems with Heckman models. First, Heckman models are based on a normality assumption, and so most event data analyses will have to transform their dependent variable in some way to account for Poisson distributions. In addition, Heckman models are reliant on correctly modeled behavior (Winship and Mare 1992). Omitting important variables produces biased results. Previous studies have used Heckman models to assess the effects of selection bias to show that selection bias does not affect the relationship between democracy and inequality across three studies (Hughes 1997). Indeed, Hug and Wisler (1998) find that modeling endogenous selection biases directly is worthwhile when selectivity is severe and the factors affecting selection (such as those listed above) are known (see also Hug 2003).

A second approach is to treat nonrandom sampling as exogenous. This means that the selection is being made outside of the media process, for example, by governmental censorship.

One approach to modeling exogenous factors is inflation models, such as hurdle models and zero-inflated Poisson (ZIP), and negative binomial (ZINB) models. Inflation models are two-stage models that combine logit and count models to predict whether events are included or excluded from the data set, explicitly model the factors that should theoretically influence the selection process, and predict the frequency of behavior (Long and Freese 2006).

Inflation models depend on context and case-specific knowledge of the nature of selection bias, which limits generalizability and challenges researchers to identify and measure all the sources of bias. Yet, treating bias as an exogenous and explicit part of the models enables scholars to further theorize and assess the selection process directly. For instance, Hill, Moore, and Mukherjee (2013) used zero-inflated probit models to show that increased media reports, Amnesty International (AI) reports, and terror attacks increase the probability that AI exaggerated torture allegations, and scholars, such as Crenshaw, Robison, and Jenkins (2014) have controlled for press freedom, total population size, and story counts (i.e., the "newshole"), to control for inflated "zeros" (see also Bagozzi et al. 2015).

A third and promising approach is to use simulation models to assess the effects of selection on the findings of studies (Imai and Yamamoto 2010). These approaches are essentially updated versions of jackknife resampling methods that draw on recent advances in computational power to estimate confidence in results (Cameron and Trivedi 2005). There are a variety of simulation approaches, but they all use existing data ranges to generate a large number of simulated data sets based on quantitative models of the nonrandom error that can be analyzed to assess the influence of nonrandomness on results. Gallop and Weschle (2015) propose a simulation-based sensitivity analysis that simulates different levels of bias in the data in order to assess the susceptibility of results to nonrandom error and the level of bias at which a hypothesis is no longer supported. This enables analysts to identify different levels of possible bias, determine robustness, and establish confidence levels. Gallop and Weschle's (2015) approach is flexible and useful because it is not limited by level of measurement, and it makes no assumption about the actual structure of the nonrandom error.

Hill and Jones (2014; see also Breiman 2001) use cross-validation and random forest methods to assess the predictive power of specific variables added to the statistical model.

Cross-validation randomly divides the data set a number of times in order to evaluate the model's ability to predict the outcome. Random forests take random selections of the available data, identify the variable that is most strongly related to the dependent variable, and then select the variables that are consistently predictive of the outcome. This approach addresses nonrandom selection by using simulated random sampling to effectively minimize the effects of selection while determining the predictive validity of each measure. For this approach to fail, the news selection process would have to be so severe that the base sample from which each simulated random sample is drawn would have to be significantly different from the "true" population. With this in mind, it is notable that the findings of Hill and Jones (2014) match the strongest findings from previous repression research using regression models (specifically the effects of conflict and democracy; see Davenport [2007]). Furthermore, this fits with our general contention that our focus should be on trying to assess and control for the effects of selection bias instead of on the impossible task of eliminating selection problems entirely.

CONCLUSIONS

This review has attempted to outline a strategy for dealing with source selection in event data and for building a new generation of event data systems and methods that will be able to assess and reduce the effect of possible selection bias in event data. The promise of harnessing the Internet and making use of new tools that have been developed for the identification and automated construction of event data is currently stymied by the inability to deal with this problem. As long as we lack a strategy for addressing the problem of possible selection bias, we will not be able to move to a higher level of analysis, assess the generalizability of our findings, and make causal assessments.

At its heart, the news selection problem stems from the fact that, with event data, there is no known universe of "real world" events that we can sample or against which we can compare. We have only a variety of partial and limited samples, some of which are more complete than others, where we know little about their randomness. Nor, given the nature of news data, is this going to be resolved in ways that have been conventionally adopted with other forms of social science data, such as random sample surveys. There is no universe of relevant events or a simple device, such as random digit dialing, that would give us a random sample of "real world" events.

What to do? We argue that we should abandon any pretense that there is an absolute base code for assessing random selection and instead adopt a relativistic strategy to assess the severity of the selection problem and ways to minimize the risks of inference errors. Instead of holding out the false hope that we will eventually have a fully random, error-free sample of events, it seems more promising to devise methods that allow us to assess how serious the problem is and ways to minimize the risks of false inferences. In this spirit, we have outlined a series of methods for constructing multisource event data and for assessing selection bias. In particular, we have summarized a series of techniques for controlling for endogenous and exogenous sources of nonrandomness and assessing the severity of selection bias.

We began with a summary of the problem, which suggests that single-source data sets are typically more vulnerable to selection bias. Although there are instances where adding

additional sources may only magnify the biases, in most cases, if there are sufficient resources, it seems advantageous to draw on multiple sources. There are no single "authoritative" data sources, such as police records, that can be used as full population samples. But there are multisource inventories that seem less vulnerable to problems. At the minimum, getting beyond inventories that contain only 20–40 percent of the events to larger, multisource inventories would provide a first step for improving the quality of event data.

Beyond this, event data would be strengthened by having better information on the problem of temporal and geographic stability in event selection. At present, we have only a few studies that provide conflicting evidence about the severity of the problem. Certainly ambitious projects, such as creating a global data system for the monitoring of atrocities and humanitarian early warning, need to address the question of temporal and geographic stability in event data. We also need more studies of the consistency of regression results that come from the analysis of different event inventories. At present, we have only a handful of such studies, without which a firmer sense of the severity of the selection problem is impossible. Yes, a single source that provides only 20 percent of the events that a larger multisource inventory provides does seem to be a risky basis for causal inference. Ultimately, the problem is that we really do not know how bad the selection problem is, so better studies are needed.

A third area of needed work involves devising additional methods for constructing multisource inventories, especially with automated or machine-learning methods. Current efforts have demonstrated the feasibility of creating such data and, with further refinements, it seems likely that methods for the automated construction of multisource inventories can be accomplished. This requires addressing hard problems, like the resolution of duplicate reports and improving coding accuracy, but in principle, these seem to be ultimately soluble. One issue that will have to be addressed is the trade-off between event detail and sparse events (i.e., simple "actor/event form/target" data). The more detailed the event attributes, the less the accuracy of our coding. So we need to be cognizant of where this trade-off should be made.

A fourth area involves devising methods for assessing and controlling for news selection. Current methods for dealing with endogenous and exogenous sources of event selection have just begun to address the many possibilities that may exist. In part this is tied to the substance of particular studies where, for example, press freedom or operational constraints of the news system are key parts of the selection process. Identifying and bringing these into the analysis seems to be the best approach for reducing false inference. Simulation methods also promise to give us a better sense of the severity of the problem and what parts of our findings can survive the challenge.

Our main message is that a relativistic approach that recognizes the inevitably limited and partial nature of our data is a healthier tack. While we often tend to fall back into the assumption that there is a single absolute standard for identifying random samples, event data is not a field where this model will apply. In fact, as some have noted, other fields of social science confront similar problems and have to devise methods that allow us to assess the risks and move on. Schrodt's (2012) admonition that event data seems to be in a situation analogous to that of survey analysis prior to the acceptance of random sampling seems apt. By adopting a strategy more attuned to the real possibilities of building stronger event data, the field may be able to progress.

REFERENCES

Ackland, Robert. 2009. "Social Network Services as Data Sources and Platforms for e-Researching Social Networks." *Social Science Computer Review* 27(4):481–92.

Almeida, Paul, and Mark I. Lichbach. 2003. "To the Internet, From the Internet: Comparative Media Coverage of Transnational Protests." *Mobilization* 8(3):249–72.

Amenta, Edwin, Near 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.

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.

Bagozzi, Benjamin E., Daniel W. Hill, Will H. Moore, and Bumba Mukherjee. 2015. "Modeling Two Types of Peace: The Zero-Inflated Ordered Probit (ZiOP) Model in Conflict Research." *Journal of Conflict Resolution* 59(4):728–52.

Barranco, Jose, and Dominique Wisler. 1999. "Validity and Systematicity of Newspaper Data in Event Analysis." *European Journal of Sociology* 15(3):301–22.

Betz, Timm. 2013. "Robust Estimation with Nonrandom Measurement Error and Weak Instruments." *Political Analysis* 21(1):86–96.

Bond, Doug, J. Craig Jenkins, Charles L. Taylor, and Kurt Schock. 1997. "Mapping Mass Political Conflict and Civil Society." *Journal of Conflict Resolution* 41(4):553–79.

Boycoff, Jules. 2006. *The Suppression of Dissent: How the State and Mass Media Squelch U.S. American Social Movements*. London: Routledge.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Chojnacki, Sven, Christian Ickler, Michael Spies, and John Wiesel. 2012. "Event Data on Armed Conflict and Security: New Perspectives, Old Challenges and Some Solutions." *International Interactions* 38(4):382–401.

Crenshaw, Edward, Kris Robison, and J. Craig Jenkins. 2014. "All the World's a Stage: Contentious Politics, Mass Media and Global Civil Society." Department of Sociology, Ohio State University, Columbus.

Danzger, M. Herbert. 1975. "Validating Conflict Data." *American Sociological Review* 40(5):570–84.

Davenport, Christian. 2007. "State Repression and Political Order." *Annual Review of Political Science* 10:1–23.

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

Davenport, Christian, and Patrick Ball. 2002. "Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977–1995." *Journal of Conflict Resolution* 46(3):427–50.

Earl, Jennifer. 2013. "Studying Online Activism: The Effects of Sampling Design on Findings." *Mobilization* 18(4):389–406.

Earl, Jennifer, and Katrina Kimport. 2011. *Digitally Enabled Social Change: Activism in the Internet Age*. Boston, MA: MIT Press.

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* 30:65–80.

Fillieule, Olivier. 1998. "'Plus ca change, moins ca change': Demonstrations in France during the Nineteen-Eighties." Pp. 199–226 in *Acts of Dissent*, edited by Dieter Rucht, Ruud Koopmans, and Freidhelm Neidhardt. Berlin: Sigma.

Gallop, Max, and Simon Weschle. 2015. "Assessing the Impact of Nonrandom Measurement Error on Inference." Department of Political Science, University of Strathclyde, UK (http://www.simonweschle.com/)

Gerner, Deborah, Phillip Schrodt, Ronald A. Francisco, and Judith T. Weddle. 1994. "The Machine Coding of Events from Regional and international Sources." *International Studies Quarterly* 38(1):91–119.

Guo, Shenyang, and Mark W. Fraser. 2014. *Propensity Score Analysis: Statistical Methods and Applications*. Thousand Oaks, CA: Sage.

Hafner-Burton, Emilie Marie, and James Ron. 2009. "Seeing Double: Human Rights Impact through Qualitative and Quantitative Eyes." *World Politics* 61(2):360–401.

Hanna, Alex. 2014. "Developing a System for the Automated Coding of Protest Event Data." *Social Science Research Network*. Retrieved October 3, 2015 (id24252132).

Heckman, James. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1):153–61.
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.
Herman, Edward S., and Noam Chomsky. 1988. *Manufacturing Consent*. New York: Pantheon.
Hill, Daniel W. Jr., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(3):661–87.
Hill, Daniel W., Will H. Moore, and Bumba Mukherjee. 2013. "Information Politics versus Organizational Incentives: When Are Amnesty International's 'Naming and Shaming' Reports Biased? 1." *International Studies Quarterly* 57(2):219–32.
Hocke, Peter. 1998. "Determining Selection Bias in Local and National Newspaper Reports on Protest Events." Pp. 131–63 in *Acts of Dissent*, edited by Dieter Rucht, Ruud Koopmans, and Friedhelm Neidhardt. Berlin: Wissenschaftszentrum Berlin für Sozialforschung.
Hug, Simon. 2003. "Selection Bias in Comparative Research: The Case of Incomplete Data Sets." *Political Analysis* 11(3):255–74.
Hug, Simon, and Dominique Wisler. 1998. "Correcting for Selection Bias in Social Movement Research." *Mobilization* 3(2):141–61.
Hughes, Marion. 1997. "Sample Selection Bias in Analyses of the Political Democracy and Income Inequality Relationship." *Social Forces* 75:1101–1117.
Imai, Kosuke, and Teppei Yamamoto. 2010. "Causal Inference and Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis." *American Journal of Political Science* 54(2):543–60.
Jenkins, J. Craig, Charles Lewis Taylor, Marianne Abbott, Thomas Maher, and Lindsey Peterson. 2014. "Global Conflict Data: Introducing the World Handbook of Political Indicators IV Data Set." Department of Sociology, Ohio State University, Columbus (www.sociology.osu.edu/people/jenkins.12)
King, Gary, and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders." *International Organization* 57(3):617–42.
Koopmans, Ruud. 2004. "Movements and Media: Selection Processes and Evolutionary Dynamics in the Public Sphere." *Theory and Society* 33(3–4):367–91.
Leetaru, Kalev, and Philip A. Schrodt. 2013. "GDELT: Global Data on Events, Locations and Tone, 1979–2012." Department of Political Science. Paper presented at the Annual Meeting of the International Studies Association, April 4, 2013, San Francisco.
Long, J. Scott, and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. College Station, TX: Stata Press.
Martin, Andrew W. 2005. "Addressing the Selection Bias in Media Coverage of Strikes: A Comparison of Mainstream and Specialty Print Media." *Research in Social Movements, Conflict and Change* 26:141–78.
McAdam, Doug, and Yang Su. 2002. "The War at Home: Antiwar Protests and Congressional Voting, 1965–1973." *American Sociological Review* 67(5):696–725.
McCarthy, John D., Clark McPhail, and Jackie Smith. 1996. "Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991." *American Sociological Review* 61(3):478–99.
McCarthy, John D., Larissa Titarenko, Clark McPhail, Patrick S. Rafail, and Boguslaw Augustyn. 2008. "Assessing Stability in the Patterns of Selection Bias in Newspaper Coverage of Protest during the Transition from Communism in Belarus." *Mobilization* 13(2):127–46.
Mueller, Carol. 1997. "International Press Coverage of East German Protest Events, 1989." *American Sociological Review* 62(5):820–32.
Myers, Daniel J., and Beth S. 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):519–43.
Oliver, Pamela E., and Gregory Maney. 2000. "Political Processes and Local Newspaper Coverage of Protest Events: From Selection Bias to Triadic Interaction." *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.
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.
Parenti, Michael. 1993. *Inventing Reality: The Politics of the News Media*. New York: St. Martin's Press.

Rohlinger, Deana A., Ben Kail, Miles Taylor, and Sarrah Conn. 2012. "Outside the Mainstream: Social Movement Organization Media Coverage in Mainstream and Partisan News Outlets." *Research in Social Movements, Conflicts and Change* 33:51–80.

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

Seybolt, Taylor B., Jay D. Aronson, and Baruch Fishhoff eds. 2013. *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*. New York: Oxford University Press.

Shellman, Stephen M. 2008. "Coding Disaggregated Intrastate Conflict: Machine Processing the Behavior of Substate Actors over Time and Space." *Political Analysis* 16(4):464–77.

Snyder, David, and William R. Kelly. 1977. "Conflict Intensity, Media Sensitivity and the Validity of Newspaper Data." *American Sociological Review* 42(1):105–23.

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

Ward, Michael D., Nils W. Metternich, Cassy L. Dorff, Max Gallop, Florian M. Hollenbach, Anna Schultz, and Simon Weschle. 2013a. "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction." *International Studies Review* 15:473–90.

Ward, Michael D., Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. 2013b. "Comparing GDELT and ICEWS Event Data." *Analysis* 21:267–97.

White, Robert. 1993. "On Measuring Political Violence: Northern Ireland, 1969 to 1980." *American Sociological Review* 58(4):575–85.

Winship, Christopher, and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18:327–50.