What Should We Do About News Selection in Event Data?

Challenges, Progress and Possible Solutions

J. Craig Jenkins
Ohio State University
Jenkins.12@osu.edu

Thomas V. Maher
University of Arizona
thomasvmaher@email.arizona.edu

ABSTRACT

What Should We Do About News Selection in Event Data?

Challenges, Progress and Possible Solutions

The prospect of making use of 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. Studies have shown that event data has significant news selection problems, especially in studies using single sources. In terms of conventional standards of representativeness, the extent of news selection and the departure of event data samples from randomness is largely unknown. We summarize recent studies of news selection and outline a series of methods for reducing the risks of possible selection bias, including techniques for generating multi-source event inventories and analytic methods for estimating and controlling for non-randomness. 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.

What Should We Do About News Selection in Event Data?

Challenges, Progress and Possible Solutions

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 contentious politics (e.g. Almeida and Lichbach 2003; Bond et al. 1997; Earl and Kimport 2011; Gerner et al. 1994; King and Lowe 2003; Shellman 2008; Schrodt 2012; Chojnacki et al. 2012; Leetaru and Schrodt 2013; Hanna 2014; Jenkins et al. 2014). But this promise confronts a major challenge in terms of our ability to say much about the representativeness of protest event data. Facing the same challenge confronted by other attempts to harness "Big Data" in the social sciences, protest event data faces major questions about its ability to meet conventional standards of representativeness and reliability. Two recent 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: "(R)esearchers 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 much more pessimistic conclusion: "(N)ewspaper 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 questions. First and foremost, how serious is the problem of news 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? Finally, are there solutions in terms of data collection procedures and analytic methods that can address the problem of news selection? In this discussion, we draw on insights from social movement studies and work done by political scientists in the study of international relations and comparative politics.

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 that is generalizable and to draw causal inferences, then knowing that 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, e.g., to predict policy change (McAdam and Su 2002), in which case it is arguable that the media constitutes the major mechanism through which elites become aware of protests, then selection may not be a problem. If, however, protest is the dependent variable, then addressing the question of random selection is a central concern.

The Problem of News Selection

At its core, the problem of news selection stems from the fact that we lack comprehensive population-level inventories of all protests (or other 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 that constitute the full population of events. The resulting bias may be small or large but the problem is we have no direct way to randomly sample an unknown universe.

How serious is the problem of news selection? One crude starting point is the proportion of protests that are covered in one source compared to another or to a multi-source 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 basecode, such as an integrated multi-source event inventory or a single authoritative source such as police records, is used for comparison. Typically, a single news source covers no more than 20 to 40 percent of the full inventory of events and often the rate is significantly lower. For example, in a study comparing Swiss newspaper reports of protest against police records, Barranco and Wisler (1999) found relatively high rates approaching 50 percent of the police reports. By contrast, Fillieule (1998) found only 2 to 3 percent of the police-reported protests were also reported in national French newspapers. In a study of local and national newspaper coverage of protests in Freiberg Germany, Hocke (1998) found that the local paper 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. At the same time, the protests reported in the national papers were the very largest events and all of those involving violence, indicating the selection dynamics operating in the national newspapers.

How much does the result depend on the goals of the news source? Some argue that state-owned media and ideological or partisan news sources have a political stake in reporting and will either ignore anti-regime 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 "copy-cat" 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 not selective ideologically but were more likely to report on events organized by groups with professional staff. Professional staff not only provided legitimacy and credibility but also a liaison for collecting information. By contrast, partisan news sources (both conservative and liberal) tend to report events of groups sharing their ideology but did not differentiate between professional and voluntary groups.

A similar picture is provided by Davenport's (2010) study of the mainstream vs. 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 by authorities against activists. This coverage made it appear that governmental institutions were in control and that dissidents were largely reactive, responding to legal controls and repression with coercive and violent dissent. In contrast, the black power movement press provided more coverage of non-contentious and non-coercive Party activities and of the police coercion that targeted activists. The picture is one of active dissidents engaged in institutional activities which are met with police repression, which then triggered non-violent

dissent. Which is the more accurate? Davenport (2010) argues that this is an invalid question and that a more accurate picture is provided by integrating these two event catalogues while remaining aware of the different perspectives of the different sources.

One study that contravenes this conclusion about types of news sources is McCarthy et al.'s (2008) study of protest reports in state-owned vs. private for-profit newspapers in Minsk, Belarus during the immediate post-Communist transition. In this study, the four individual newspapers each reported about 30 percent of the protests found in police records and, when combined together, covered around 38 percent of the police reported events. Comparing state-owned vs. 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; Oliver and Myers 1999; Oliver and Maney 2000; Myers and Caniglia 2004; Swank 2000). While the size of the "newshole" might be somewhat elastic, the ultimate size of the available "print space" is a constraint. In regression models, one way to control for this "newshole" is to statistically control for a count of the number of news stories, which could be used in a hurdle model or as a source of negative selection in a zero-inflated poisson (ZIP) regression (Long and Freese 2006). In a cross-national time-series study of protests and terror attacks, Crenshaw et al. (2014) used a ZINB model, finding that the number

of Reuters newswire stories mentioning a particular country created excess "zeros" for protest and terror attacks.

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 (Bagdikian 2000; Parenti 1993; Herman and Chomsky 1988; Boycoff 2006). 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. 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. Further, 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. Fitting with the ongoing thematic agenda of the newspaper is important. In a similar vein, Oliver and Myers (1999) found that public events with business sponsors were more likely to be covered, typically in the "business" section of the paper.

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. Insofar as prepackaged news stories delivered by wire services are replacing locally written stories, news wires would appear to be more comprehensive sources of

protest reports, a contention supported by Strawn's (2008) evidence on national news agency vs. regional newspaper reportage in Mexico.

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 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). Reporters often turn to government officials for information on developing events. Events that lend themselves to such key informants are more likely to be reported. As newspapers downsize their reporting staff and rely more heavily on centralized newswire services for content, the coverage of certain kinds of events, among them social movement activities, are less likely to be covered.

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," i.e. 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. In specific, 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; McCarthy et al. 1996; Mueller 1997a; Oliver and Myers 1999; Fillieule 1998; McCarthy et al. 2008; Herkenrath and Knoll 2011);

(2) the geographic distance between the event and the media source, especially its reporting market and audience (Barranco and Wisler 1999; Hocke 1998; McCarthy et al. 1996; Mueller 1997a; Fillieule 1998; Strawn 2008; Herkenrath and Knoll 2011);

(3) the extraordinariness of the event in terms of unruliness, arrests, violence, the presence of counter-demonstrators, and the flamboyance of events (Snyder and Kelly 1977; Barranco and Wisler 1999; Mueller 1997a; Oliver and Myers 1999; Myers and Caniglia 2004);

(4) protest that fits with the ideological stance and the thematic coverage priorities of the media source (Barranco and Wisler 1999; Mueller 1997a; Oliver and Maney 2000; Rohlinger et al. 2012; Andrews and Caren 2010; but see McCarthy et al. 2008);

(5) the existence of favorable media interest (McCarthy et al. 1996; Fillieule 1998; Oliver and Maney 2002; Rohlinger et al. 2012);

(6) the legitimacy and professionalism of the event sponsor, including the presence of celebrities and other sources of legitimacy (Snyder and Kelly 1977; Oliver and Maney 2002; McCarthy et al. 2008; Andrews and Caren 2010; Rohlinger et al. 2012).

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 the power of partisan allies does not influence selection.

These studies have led to the general conclusion that multi-source 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 multi-source 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 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, i.e. the basecode is seen as constituting the full population or universe of events? Or is this a relative inference problem in which one is attempting to identify sources of 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 actually is important to judging what to do with limited and partial inventories and for conceptualizing the larger strategy for dealing with selection bias. Is one to toss out these limited and partial datasets? Or is the best approach to devise ways to assess their contributions in light of their limitations?

In this regard, it is well to keep in mind Oliver and Myers' (1999: 48) conclusion 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 regards 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 also 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 datasets 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. Additionally we need to know what the procedures are for identifying duplicate reports of the same underlying event. Adding additional news sources increases the likelihood of multiple reports about the same event. How is that "same event" identified? Does one use the additional information provided in multiple reports to identify event characteristics? Finally, we need to recognize that "authoritatitive" non-media sources such as police records (Oliver and Myers 1999), the state (Scott 2000), and NGOs (Hafner-Burton and Ron 2009) are likely have their own selection biases, which also need to be taken into account.

Does News Selection Affect Study Results?

A second more precise way of answering the question about the severity of the news selection problem 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 news sources?

Unfortunately only a few studies have actually addressed this question. In the Davenport and Ball (2002) study, they 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 1-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* vs. a larger multi-source 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 only relevant 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 might actually be relatively limited.

A third study addresses the question of temporal stability. In their study of post-Communist 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 different time periods (i.e. the post-Petrestroika 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" (142). Further, they note their findings about protest size are consistent in other country studies (e.g. Switzerland, the U.S., 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. He 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" (White 1993: 583).

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* vs. those provided by a multi-source 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 (but not 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 dataset vs. *NYT* and *NYT-Washington Post* coverage. While most of the effects are statistically significant in all three equations (black population size, location in the south, proportion foreign born, history of disorder), 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* dataset [p = .102]) and another (black median income) is statistically significant only in the combined and *NYT* datasets. Further, they argue that the sizes of the coefficients relative to their standard errors are much larger in the *NYT* dataset, 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 a relative deprivation process behind the urban riots but this is only one of six statistically significant factors. Overall, their main point is clear: Multi-source 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 such as *Factiva* and *Lexis Nexis* coupled with the development of computational tools for coding large amounts of electronic text into event data promises the possibility of constructing multi-source event data sets (Bond, et al. 1997; Gerner et al. 1994; King and Lowe 2003; Shellman 2008; Schrodt 2012; Leetaru and Schrodt 2013; Hanna 2014; Jenkins et al. 2014). Despite this promise, this development confronts several challenges. First, analysts have found that online

news archives are often unstable with additions and deletions of news stories to fit the space limits of the archive and to maintain timeliness of news coverage (Ortiz et al. 2005). Over time, news archives have tended to grow, which improves their coverage but creates temporal inconsistency.

A second problem is coding reliability. Automated coding has great advantages in terms of processing large volumes of text but actually the main argument for machine coding is transparency and consistency. Computer errors are transparent (unlike human coding which has an inevitable black box quality to it), which means the computer code can be corrected and the text analysis rerun. This does not, however, necessarily insure high reliability. In a comparison of machine vs. human coding, King and Lowe (2003) found that machine coding using the VRA KnowledgeManager parser was comparable in accuracy to human coding—in some cases as low as 25% to 50% for the detailed event types (e.g. political graffiti) and up to 55 to 70% for the more generic cue category events (e.g. protest demonstrations). In general, the simpler the coding scheme, the greater the accuracy. Before one concludes that such coding accuracy is unacceptable, it is well to realize that most studies of human coding show similar or lower rates of reliability for complex codes (see Mikhaylov et al. 2012; Ruggeri et al. 2011). As Schrodt (2012: 555) concludes, "the sustained decision making required for human coding presents almost a perfect storm for inducing fatigue, inattention, and a tendency to use heuristic shortcuts…. The human brain was simply never intended for the tasks we impose on coders."

A third problem is identifying and resolving information on multiple reports of the same underlying event. 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 multi-source event database that makes use of Googlenews.com to create "near real-time" daily updates of violent and other events along with other conflict indicators (Leetaru and Schrodt 2013). Several analysts using GDELT for humanitarian early warning have noted that GDELT does not currently clean or mark for duplicate reports. As a result, a naïve 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). The point is that duplicate reports need to be identified and resolved. In the process, analysts need to establish clear minimum criteria for defining a match on the same event and decide how to deal with additional information on event features. 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 websites, blogs and the like, which may provide event summaries. In a unique study, Almeida and Lichbach (2003) compared the protest counts on activist websites 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 websites are more complete than any other source, reporting almost half of a larger multi-source inventory, and are less selective

with regards to 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 SMO protests located in Seattle, the international news archive *Lexis Nexis* reported the greatest number of events and, for national events outside of Seattle, the *New York Times* reported the least. However, not all activist websites are equally valuable. To build this database, they consulted 20 websites. Websites with a special news section were the most valuable, providing event chronologies, archiving messages and eyewitness reports, and maintaining electronic hyper-links to news articles elsewhere. A significant challenge was developing procedures to confirm that reported events actually occurred by, e.g. comparing web reports against local media coverage. Of course, not all protest campaigns will have websites that are constantly updated and maintained.

A final issue in making use of the Internet is 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 websites. 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" websites 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 websites (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 1000+ results, which concatenated generates 6,000 to 14,000 results. Once cleaned for duplicates, this list of websites then constitutes the sampling frame of public websites 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 to 40% more websites 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 websites" 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.

How About Analytic Fixes?

A final question is whether news selection can be treated through analytic methods. Scholars have developed a number of approaches for modeling non-random error which in principle can be applied to event data. The most common is to treat this error as endogenous. This makes sense in that 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 this process. But the media is also a separate actor, and so it is important to delineate clearly the roles that the state, movements, and media play in the process.

The most common approach to modeling endogenous non-random sampling error is Heckman models (Heckman 1979; see also Maddala 1983; 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. Additionally, 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 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 (1998; see also Hug 2003, 2010).

A second approach is to treat non-random sampling as exogenous. This means that the selection is being made outside of the media process, e.g. by governmental censorship.
One approach to modeling exogenous factors is inflation models, such as hurdle models, zero-inflated poisson (ZIP), and zero-inflated negative binomial (ZINB). These are basically two stage models that combine logit and count models to predict whether events are included or

excluded from the dataset (i.e. coded as "zero" when in fact there are events), 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 used zero inflated probit models to show that increased media reports, AI reports, and terror attacks increase the probability that Amnesty exaggerated torture allegations (2013), and scholars, such as Anthony and Crenshaw (2014), have used ZINB models to model the factors, like the total flow of news stories (i.e. the "newshole"), that influence the omission of events from the analysis (see also: Bagozzi 2015; Bagozzi et. al 2015). Crenshaw et al. (2014) control for press freedom and total size of the population as well as the "newshole," making the argument that these factors create artificial or inflated "zeros" which are best controlled exogenously.

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; Hill and Jones 2014; Gallop and Weschle 2015). These approaches are essentially updated versions of jackknife resampling methods (Cameron and Trivedi 2005) that draw on recent advances in computational power to estimate confidence in results. There are a variety of simulation approaches, but they all use existing data ranges to generate a large number of simulated datasets based on quantitative models of the non-random error that can be analyzed to assess the impact of non-randomness 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 non-random error and the level of bias at which a hypothesis is no longer supported. The basic idea is to identify different levels of possible bias and see if results will hold up, thus establishing a level of confidence in results. Gallop and Weschle's 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 non-random error.

Where Gallop and Weschle use simulation models to determine the effects on the outcome, Hill and Jones use cross-validation and random forest methods to assess the predictive power of specific variables added to the statistical model (2014; Breiman 2001). Cross-validation randomly divides the data set a number of times in order to evaluate the models' ability to predict the outcome. Random forests take random selections of the available data, identifies the variable that is most strongly related to the dependent variable, and then selects the variables that are consistently predictive of the outcome. This approach addresses non-random 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 Hill and Jones' (2014) findings match the strongest findings from previous repression research using regression models (specifically the effects of conflict and democracy; see Davenport [2007]). Further, this fits with our general contention that our focus should be trying to assess and control for the effects of selection bias instead of focusing on the impossible task of eliminating selection problems entirely.

Conclusions

This review has attempted to outline a strategy for dealing with event selection in news data and for building a new generation of event data systems and methods that will be able to assess and reduce the impact 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. So 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 news data, there is no known universe of "real world" events which we can sample or against which we can compare. All we have are 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? The argument we have advanced is to abandon the idea of an absolute method for assessing random selection and instead develop a relativistic strategy that will provide ways of assessing the extent of the 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 recognize the impossibility of that notion and to devise methods that allow us to assess how serious the problem is and what we can do to

minimize the risks of making false generalizations and inferences. In this spirit, we have outlined a series of methods for constructing multi-source event data inventories and for assessing the risks of selection bias. In particular, we have summarized a series of techniques for identifying and controlling for endogenous and exogenous sources of non-randomness and for assessing whether our results have serious inference problems.

We began with a summary of the problem, which suggests that single source datasets 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, like police records, that can be used as full population samples. These too contain potential bias. But there are multi-source inventories that seem less vulnerable to problems. At the minimum, getting beyond inventories that contain only 20 to 40 percent of the events to larger, multi-source 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, like creating a global data system for the monitoring of atrocities and other events relevant to humanitarian early warning, confront a major question about how representative our event data measures are. 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 multi-source

inventory might provide 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 where we need work is devising additional methods for constructing multi-source 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 multi-source 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.[1] One issue that will have to be addressed is the tradeoff between event detail and sparse events (i.e. simple "actor/event form/target" data). The more detailed the event attributes that are collected, the less the accuracy of our coding. So we need to be cognizant of where this tradeoff should be made.

A fourth area for further work is devising more methods for assessing and controlling for news selection. Current methods for dealing with endogenous and exogeneous 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, e.g. 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 our vulnerability to 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 major message is that a relativistic approach that recognizes the inevitably limited and partial nature of our data and knowledge is a healthier tack. While we often tend to fall back

[1] Indeed, one new event data system –the Phoenix Data Project –appears to be working on these problems by building in steps to reduce false positives and noise in the data (http://phoenixdata.org/faq).

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 others 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* 27:481-492.

Almeida, Paul and Mark I. Lichbach. 2003. "To the Internet, From the Internet: Comparative Media Coverage of Transnational Protests." *Mobilization* 8:249-272.

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* 74:636-656.

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

Anthony, Robert M. and Edward M. Crenshaw. 2014. "City Size and Political Contention: The Role of Primate Cities in Democratization." *International Journal of Sociology* 44(4):7–33.

Bagdikian, Ben S. 2000. *The Media Monopoly*. Boston, MA: Beacon.

Bagozzi, Benjamin E. 2015. "The Baseline-Inflated Multinomial Logit Model for International Relations Research." *Conflict Management and Peace Science* 13:1-24.

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* 15:301-22.

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

Bond, Doug, J. Craig Jenkins, Charles L. Taylor and Kurt Schock. 1997. "Mapping Mass Political Conflict and Civil Society." *Journal of Conflict Resolution* 41:553-579.

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 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* 38:382-401.

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

Danzger, M. Herbert. 1975. "Validating Conflict Data." *American Sociological Review* 40:570-584.

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.* N.Y.: 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* 46:427-450.

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

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,* ed. Dieter Rucht, Ruud Koopmans and Freidhelm Neidhardt. Berlin: Sigma.

Gallop, Max and Simon Weschle. 2015. "Assessing the Impact of Nonrandom Measurement Error on Inference." Dept. 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* 38:91-119.

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

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

Hanna, Alex. 2014. "Developing a System for the Automated Coding of Protest Event Data." *Social Science Research Network* (id24252132) Downloaded 10/3/2015.

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.* N.Y.: Pantheon.

Hill, Daniel W. Jr. and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108:661-687.

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. 131-163 in *Acts of Dissent,* ed. Dieter Rucht, Ruud Koopmans and Friedhelm Neidhardt. Berlin: Wizzenschaftzentrum Berlin Fur Sozialforschung.

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 Dataset." Dept. of Sociology, Ohio State University: Columbus OH. (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* 57:617-642.

Koopmans, Ruud. 2004. "Movements and Media: Selection Processes and Evolutionary Dynamics in the Public Sphere." *Theory and Society* 33:367-391.

Leetaru, Kalev and Philip A. Schrodt. 2013. "GDELT: Global Data on Events, Locations and Tone, 1979-2012." Dept. of Political Science, Presented at the Annual meetings of the International Studies Association, April 4, 2013, San Francisco, CA.

Long, J. Scott and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata.* College Station, TX: Stata Press.

Maddala, Gangadharrao S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University 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* 26:141-178.

McAdam, Doug and Yang Su. 2002. "The War at Home: Antiwar Protests and Congressional Voting, 1965-1973." *American Sociological Review* 67:696-725.

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* 61:478-499.

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

Mikhaylov, Slava, Michael Laver, and Kenneth Benoit. 2012. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20:78-91.

Mueller, Carol. 1997a. "International Press Coverage of East German Protest Events, 1989." *American Sociological Review* 62:820-832.

Mueller, Carol. 1997b. "Media Measurement Models of Protest Event Data." *Mobilization* 2:165-184.

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* 69:519-543.

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* 106:463-505.

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* 105:38-87.

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

Parenti, Michael. 1993. *Inventing Reality: The Politics of the News Media.* N.Y. St. Martin's.

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.

Ruggeri, Andrew, Theodra-Ismene Gizelis, and Han Dorussen. 2011. "Events Data as Bismarck's Sausages? Intercoder Reliability, Coders' Selection and Data Quality." *International Interactions* 37: 349-361.

Schrodt, Philip A. 2012. "Precedents, Progress and Prospects in Political Event Data." *International Interactions* 38:546-569.

Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, CT: Yale 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):464-477.

Snyder, David and William R. Kelly. 1977. "Conflict Intensity, Media Sensitivity and the Validity of Newspaper Data." *American Sociological Review* 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.

Swank, Eric. 2000. "In Newspapers We Trust? Assessing the Credibility of News Sources that Cover Protest Campaigns." *Research in Social Movements, Conflicts and Change* 22:27-52.

Ward et. al

White, Robert. 1993. "On Measuring Political Violence: Northern Ireland, 1969 to 1980." *American Sociological Review* 58:575-585.

Winship, Christopher, and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology 18*:327-350.

Zeitzoff