A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data *

Peter F. Nardulli,[a] Scott L. Althaus,[b] and Matthew Hayes[c]

[a] Cline Center for Democracy and Department of Political Science
University of Illinois Urbana-Champaign

[b] Cline Center for Democracy, Departments of Political Science and Communication
University of Illinois Urbana-Champaign

[c] Department of Political Science
Indiana University Bloomington

Forthcoming in Sociological Methodology

Abstract

"Big data" in the form of unstructured text poses challenges and opportunities to social scientists committed to advancing research frontiers. Because machine-based and human-centric approaches to content analysis have different strengths for extracting information from unstructured text, we argue for a collaborative, hybrid approach that combines their comparative advantages. The notion of a progressive supervised-learning approach that combines data science techniques and human coders is developed and illustrated using the Social, Political and Economic Event Database (SPEED) project's Societal Stability Protocol (SSP). SPEED's rich event data on civil strife reveals that conventional machine-based approaches for generating event data miss a great deal of within-category variance, while conventional human-based efforts to categorize periods of civil war or political instability routinely mis-specify periods of calm and unrest. To demonstrate the potential of hybrid data collection methods, SPEED data on event intensities and origins are used to trace the changing role of political, socio-economic and socio-cultural factors in generating global civil strife in the post-World War II era.

*The authors acknowledge the contributions of Loretta Auvil and Tom Redman, who provided technical expertise on the software and hardware aspects of SPEED; Ajay Singh and Gabriel Rodriguez, who trained and managed our software operators; as well as Joseph Bajjalieh and Michael Martin, who made important contributions to data management and analysis.

1. The Continuing Challenge of Unstructured Data

The “Big Data” revolution has enhanced the ability of scholars to create useful knowledge out of structured data like ordered numbers and unstructured data like text or images. This notwithstanding, fully automated processing of unstructured data still yields less sophisticated output than human analysis of texts and images. Yet because human analysis does not scale well, this traditional solution to unstructured data analysis cannot exploit the explosion of information generated by the data revolution. Because so much information that is of interest to social scientists is embedded in unstructured data (Franzosi 2004; Grimmer and Stewart 2013), social researchers must find a way to leverage developments in data science if they are to advance social science knowledge and keep pace with other disciplines.

Both the challenges and opportunities posed by “Big Data” can be seen in research on civil strife, which has become an increasingly important topic for conflict and development scholars. Data on civil strife events are traditionally harvested from news reports. Recent advances in information technologies have led to the proliferation of digitized news sources and prodigious amounts of digitized news content. Technological innovations have also enhanced the ability of researchers to access this diverse array of available content. These developments make it possible to mitigate what civil strife researchers have long known to be an important validity threat to media-based data: selection biases that may distort the true distribution of events (e.g., Woolley 2000; Althaus et al. 2011; Danzger 1975; Franzosi 1987; Jackman and Boyd 1979). Unfortunately, conventional approaches to content analysis require analysts to make a trade-off that limits the methodological benefits of diverse news sources: they must choose between the number of documents to be analyzed and the richness of the data to be extracted. Although human-based coding has long been the norm for content analysis, the advent of “Big Data” has made its limits increasingly apparent.

This article introduces a hybrid model that leverages the strengths of both machine-based and human-centric approaches to content analysis. Its basic thesis is that, until fully automated approaches can match the flexibility and contextual richness of human coding, the best option for generating near-term advances in social science research lies in hybrid systems that rely on both machines and humans for extracting information from unstructured texts. The utility of hybrid approaches has been demonstrated in document clustering applications (Grimmer and King 2011), and has been recognized by such technologically advanced companies as Google, IBM and Apple (Lohr 2013). Hybrid systems are a key part of the international email-monitoring system developed by the National Security Agency (Savage 2013). This paper illustrates the value of such hybrid approaches for social research using data generated by the Social, Political and Economic Event Database (SPEED) project. SPEED’s information base contains over 100 million news reports drawn from every country in the world and spans the period from World War II to the present; its Societal Stability Protocol (SSP) generates civil strife data. The enormous volume of unstructured data embedded in these news reports is transformed into quantitative data by a structured set of workflows, developed in collaboration with a team of data scientists, that yield more nuanced codings than automated systems can provide at scales that were previously impossible for human coders.

The next section introduces two conventional approaches to information extraction and illustrates how they have been employed to study civil strife. The following section outlines SPEED's hybrid approach; the fourth section examines its value-added by contrasting SPEED data with existing civil strife data. The fifth section illustrates the potential for creating new frontiers in this field; the last section concludes.

2. Alternative Approaches to Information Extraction

The systematic analysis of textual content by human coders has been a standard method within social science research since at least World War II (Krippendorff 2004; Neuendorf 2002). For almost as long, social scientists have experimented with using computers to emulate human-coded methods at larger scales, higher speeds, and with more precision (e.g., Holsti 1964; Stone 1962). A half-century later, social scientists continue to experiment with the computational analysis of textual data (e.g., Monroe and Schrodt 2008; Franzosi, De Fazio, and Vicari 2012). Yet for all the technological sophistication of today's machine-driven methods, the computational analysis of unstructured data is still far from matching the interpretive nuance and sophistication of human-based approaches.

2.1 *Machine-based and Human-centric Approaches*

The limitations of machine-based approaches can best be understood by distinguishing between “manifest” and “latent” content in textual data (Berelson 1952). Manifest content is observable and can be recognized without making a subjective judgment about its presence: it is there, and obviously so, or it is not. Examples of manifest content include the number of words in a newspaper story, whether it appears on the front page, and whether the word “protest” appears in the text. In contrast, latent content resides in symbolic patterns within in a text that must be holistically interpreted by a qualified analyst (Potter and Levine-Donnerstein 1999). Examples of latent content include identifying the text’s topical focus, whether it depicts the government in a positive light, and whether a quoted source is arguing against a course of action. These symbolic patterns sometimes reside in the manifest content of a text. But proper interpretation of latent content often requires the analyst to draw from contextual knowledge not contained within the text.

Human- and machine-based approaches can be compared along an array of dimensions (for recent discussions, see Quinn et al. 2010; Young and Soroka 2012), but two are particularly important: their efficiency in analyzing large amounts of text and their ability to capture latent content. Machines are unquestionably better at analyzing manifest content more quickly, consistently and accurately than humans. Computers also trump humans at naive inference and straightforward forms of deductive reasoning. Thus, machine-based approaches are preferred when information extraction involves simple concepts, is context-free, or when the context required for identifying latent content can be derived entirely from the text’s manifest content (e.g., Liu 2011; Witten, Frank, and

Hall 2011). Examples include translating keywords into numerical scores for sentiment analysis; implementing rule sets that make simple queries about a text; and deriving patterns solely from manifest content (e.g., What are the most common verbs in sentences containing the word ‘demonstration?’). Thus, machines outperform humans with respect to such tasks as named entity extraction, topically classifying documents based on word co-occurrences, and determining whether different documents contain similar content.

Human-centric approaches can perform well in extracting latent content, particularly when coding judgments are context-dependent or involve complex features of text or multiple dimensions of content. The comparative advantages of trained humans derive from their capacity to draw from contextual knowledge not embedded in the text, to deal with different patterns of language usage, and to employ inductive reasoning to disambiguate textual references and meanings (Potter and Levine-Donnerstein 1999). Unlike computers, humans usually have little trouble determining which of several named persons any given “she” refers to; whether a date refers to the day of an attack or of a protest; and whether a quoted source is providing self-serving information. Moreover, humans can make complex judgments required to deal with ambiguity (e.g., understanding that a reference to “their concerns” elaborates upon a passage in the previous sentence) and determine how to apply contextual information across multiple levels of analysis (e.g., whether the headline is useful for interpreting an ambiguous phrase). Humans remain superior to machines when the task requires inductive reasoning to spot a complex pattern, or to decide how that complex pattern should be coded.

The extraordinary recent advances in computing capacity, machine learning, and natural language processing (NLP) have improved the ability of machines to emulate human processing of latent content. However, the efficiencies of machine-based approaches have most consistently been realized using the “bag of words” approach, which considers only patterns of word co-occurrence and proximity and ignores a text’s narrative structure (e.g., Evans et al. 2007; Quinn et al. 2010). More complex forms of NLP use syntactical structure to derive meaning (e.g., van Atteveldt et al. 2008; Sudhahar et al. 2013). But because they are so computationally intensive and require texts that strictly adhere to formal grammatical rules, they are difficult to employ at scale and on casually-structured texts like news reports. A more general concern with machine-based coding is that their results are rarely validated against human judgments. Sometimes computational methods deliver results similar to those of humans (e.g., King and Lowe 2003; Soroka 2012), and sometimes not (Conway 2006). But mainly the comparisons are never attempted both because they are time-consuming and because many machine-based applications – like document clustering – have no obvious human-coding counterpart.

One machine-based approach that allows for testing both the reliability and validity of automated coding processes is a supervised-learning approach (for recent reviews, see Evans et al. 2007; Hillard, Purpura, and Wilkerson 2008; Witten, Frank, and Hall 2011). This approach employs two sets of human codings: a “training set” of documents used to teach a machine system to optimally match human codings, and a “test set” of documents used to test the accuracy of the machine’s outputs. The investment required to obtain these human-coded training and test sets can be substantial – they often consist of thousands, and ideally tens of thousands, of coded documents.

Once a machine learning system demonstrates that it can reasonably reproduce human judgments it is released to operate independently without further interaction with humans. However, applying the machine system to a different information source, or checking the system's accuracy over time, requires new sets of human-coded data. Conventional supervised learning systems therefore achieve optimal efficiencies over human coding only when the project parameters are invariant, the scale of the coding project is large, and the coding task straightforward.

Supervised learning systems aside, the extent to which machines can reliably emulate human judgments remains an open question. One conclusion is clear: machine coding is no simple substitute for human coding. Earlier generations of researchers were warned against placing unwarranted faith in methodological shortcuts like factor analysis (e.g., Armstrong 1967) and stepwise regression (e.g., Thompson 1995). The same cautions apply to automated processing of unstructured data, which presents no magic fix for the challenges in this field (Grimmer and Stewart 2013). Research on civil strife illustrates this point.

2.2 *Existing Approaches to Civil Strife Event Data*

The contemporary importance of civil strife is well-articulated by Kahl (2006, 1):

> Civil strife in the developing world represents perhaps the greatest international security challenge of the early twenty-first century. Three-quarters of all wars since 1945 have been within countries rather than between them…Wars and other violent conflicts have killed some 40 million people

> since 1945, and as many people have died as a result of civil strife since 1980 as were killed in the First World War.

In addition to the post-WWII shift from international to domestic conflict (e.g., Lacina and Gleditsch 2005; Themnér and Wallensteen 2011), scholars have noted that, while civil wars have declined since 1992, other types of civil strife such as state repression, one-sided violence, and non-violent protests have increased (Bernauer and Gleditsch 2012, 377; Urdal and Hoelscher 2012; Themner and Wallensteen 2011). These changes in the makeup of conflict have spurred a renewed interest in generating civil strife event data . These efforts have included both human-centric and machine-based approaches.

The primary dependent variables in the study of civil strife derive from either *episodic* or *discrete* event data (Schrodt 2012). As illustrated in the first column of Table 1, episodic event data like that produced by the Uppsala Conflict Data Program/ Peace Research Institute of Oslo (http://www.prio.org/Data/Armed-Conflict/UCDP-PRIO/) and the Correlates of War (COW) (http://www.correlatesofwar.org/datasets.htm) project aim to capture a series of related happenings that unfold over a relatively long period of time (weeks, months or years). Episodic data typically code a given country as being in a state of civil war (or not) during a given year; these judgments are made holistically by researchers drawing from a range of textual sources. For example, a minimum number of battlefield deaths in a country for a year might be used as a threshold for determining the existence of a civil war. Episodic data are normally operationalized as dummy variables.

TABLE 1

In contrast to episodic data, discrete event data describe specific happenings – an attack, a boycott, an arrest– that unfold over a relatively short period of hours or days. Discrete event data are normally drawn from news reports and analyzed as event counts. Schrodt notes that discrete event data can be sparse or rich. Sparse event data (second column in Table 1) typically indicate whether a type of event occurred on a given date in a given country, without providing much additional detail. Most sparse event data are, therefore, temporally precise but geographically vague. The most important sparse event datasets are fully-automated efforts like the Conflict and Media Event Observations project, or CAMEO (Gerner et al. 2002), and the Integrated Data for Events Analysis project, or IDEA (Bond et al. 2003), which is a commercial enterprise that charges for access to its data.[1] While each project required extensive upfront human investment, both are machine-based systems for extracting event data using news reports drawn principally from Reuters. Both use an automated parser that analyzes the sematic structure of independent clauses within sentences for actions that fit their typologies.

While these automated systems can quickly process huge amounts of text, the event data they generate captures little specific information on who did what to whom, or when, where and how it was done. Moreover, they are limited to the information included in a particular independent clause. Thus, in addition to losing nuanced and context-specific meanings, they also cannot capture information that is expressed in more complex semantic structures (e.g., paragraphs or narratives).

Unlike sparse event data, rich event data – like the Armed Conflict Location and Event Dataset (ACLED), the Global Terrorism Database (GTD), Social Conflict in Africa Database (SCAD), and (Worldwide Incidents Tracking System (WITS) (citations and a more comprehensive listing and overview are provided in Schrodt 2012) – contain more extensive and precise information on such things as the event's actors, intensity, location and date (column 3 of Table 1). This makes it possible to resolve these events to the city level, which means that rich event data tend to be both geographically and temporally precise. But this precision and contextual richness comes at a cost: while sparse event data collection is usually highly automated, rich event data are normally collected using human-centric approaches. To illustrate the general approach taken by rich event data projects, we examine the Armed Conflict Location and Events Dataset project, or ACLED (Raleigh et al. 2010), and the Global Terrorism Database, or GTD (National Consortium for the Study of Terrorism and Responses to Terrorism 2012). Both are human-centered content analysis projects that do not employ machine-learning techniques.

ACLED's greatest strengths are its rich diversity of sources,[2] and the fact that it captures data on actors, dates and locations. It has a "Notes" field containing various event-specific details. ACLED also has three important limitations: restricted temporal and spatial coverage, focusing mostly on African countries and recording events occurring after 1996; limited scope, considering only events that occurred within civil wars (as defined by the UCDP/PRIO dataset); and low levels of intercoder reliability (Eck 2012).

The strengths of the GTD project are its broad scope, its rich information base, and the amount of event-specific data it collects. GTD data begins in 1970 and has a global reach. GTD's historical archive was compiled by a commercial service; the 1998-

2008 data were derived from over 3.5 million news articles and 25,000 news sources. The event-specific data collected includes information on actors as well as intensity indicators. GTD's most significant limitation is its event ontology, which includes only terrorist acts. By so limiting its event ontology, GTD ignores other important forms of civil unrest and state repression. This means that analysts using GTD data are limited in their ability to examine the contextual setting within which terrorism unfolds.

In sum, it is clear that conventional data collection strategies are highly limited in their ability to advance our understanding of civil strife. Automated approaches can efficiently document whether strife events occur across large spans of space and time, but they provide limited and inconsistent information about event-specific details. Human-centric approaches provide this rich contextual information, but only for applications with relatively narrow topical, geographical and temporal horizons.

3. The SPEED Project: A Progressive Supervised-learning Approach

Dissatisfaction with the current research paradigm, together with the increasing importance and changing make-up of civil conflict, has generated calls for capturing rich event data involving a broader set of event types and actors (Eck 2012; Chojnacki et al. 2012; Salehyan et al. 2012; Raleigh 2012; Urdal and Hoelscher 2012). These calls are rooted in the need for greater flexibility and precision; the SPEED project was designed to address these needs. SPEED uses a hybrid workflow system that extracts rich event data at the city-day level from the full text of a diverse set of news sources. This hybrid system consists of both automated components and human-centric components. Together they form what we call a *progressive supervised-learning system*. In this system, human coders are presented with input data that have been automatically pre-processed and classified. Humans perform only the most difficult coding decisions, leaving the more mundane work to automated processes. Combining wide-but-shallow machine capabilities with deep-but-narrow human capabilities leverages the advantages of each while limiting their liabilities.

3.1 *Hybrid Workflow Structure*

SPEED's hybrid approach integrates machine and human components in workflows that: (1) assemble repositories of news documents; (2) classify and preprocess those documents; and (3) employ technology-enhanced humans to extract structured data. Figure 1 provides a simplified view of SPEED's workflows. The circular components of the figure pertain to databases; the triangular components refer to human activity; and the square components denote automated processes. SPEED's workflows constitute a supervised-learning system because, although unsupervised applications and humans play key roles at various stages, its initial computational algorithms were derived from human-generated training data. SPEED's approach is progressive because the design of its workflows generates human feedback that updates the system's algorithms. This provides the means to enhance the role of automation as more training data and sophisticated machine learning techniques become available.

FIGURE 1

Central to SPEED's ability to advance social research is the information base from which it draws. An interdisciplinary team spent several years building an historical news repository that draws from tens of millions of news reports carried in the *New York Times* (NYT), the Foreign Broadcast Information Service (FBIS) and the BBC's Summary of World Broadcasts (SWB) from World War II to the present. In addition to the millions of historical news reports from these sources, SPEED's contemporary news archive is continuously updated by a Heretrix web crawling system that draws daily from over 800 global news websites.[3]

SPEED's workflow begins with several stages of machine-based processing that occur before humans access the news reports (Figure 1). The first wave of pre-processing involves document classification. News stories routinely provide information on many topics that form the core of scholarly interests across a variety of domains, which is why they are a potentially valuable source of structured data. Central to realizing this potential is identifying those with information relevant to a particular research focus. Fortunately, sophisticated supervised-learning techniques have been developed to automate textual classification.[4] Implementing them, however, involves assembling a set of training/test documents – examples of articles reflecting the semantic structure of documents that are, and are not, of interest.

SPEED uses a Naïve Bayes classifier that was initially developed from a test/train set of 33,000 manually classified articles indicating the presence or absence of civil strife events.[5] Repeated tests of the classification algorithm eventually yielded a final model that correctly discarded four in five of the NYT documents and around half of SWB documents as irrelevant – obviating the need for human review.[6] Just as noteworthy is that, when the automated classifier was applied to the set of 33,000 manually classified documents, it identified many more relevant articles than the original 1,600 classified by human coders as containing civil strife information. Later re-analysis by human coders confirmed that about 1,200 of these additional articles contained relevant events. The fact that the automatic classifier identified about 75% more relevant documents than humans demonstrates that it enhances both accuracy and efficiency. As a result it far outperforms human-centric approaches to document identification.

Once the documents are classified, irrelevant documents are sent to a discard bin while relevant documents are subjected to a second wave of text analytics using Natural Language Processing (NLP) techniques. This wave extracts the names of people, locations, and organizations mentioned in the text. SPEED currently uses Apache OpenNLP to conduct the tokenization, sentence segmentation, and part-of-speech tagging required for named entity extraction. All extracted location names are then passed to a geolocation engine that automatically assigns latitude and longitude coordinates using the GeoNames geographical database (www.geonames.org), which is a deceivingly complicated task. Every location entity receives a confidence score based on the amount of evidence in a document that reduces location ambiguity. For instance, if a document contains both "Paris" and "France" then mapping this to "Paris, France" will get a higher score for "Paris". However, if it has "Paris" and "Illinois" then mapping to "Paris, Illinois" will have a higher score for "Paris".

Classified and pre-processed documents are placed into a SOLR index that stores the text of relevant news articles along with associated metadata, extracted entities, and geolocation information. This SOLR index serves as a document store, from which news articles can be channeled into analysis queues as needed (see Figure 1). Analysis queues are subsets of the document store created to extract data for specific research projects using information extracted during the preprocessing stage (e.g., dates, places, names). To generate event data from these subsets of documents, analysis queues are paired with a protocol. Protocols are electronic documents that contain structured question sets that define the information to be humanly extracted. Human coders access protocols through a web interface that integrates the protocol with a queued document.[7] Human coders are presented only those documents that have been machine-classified as relevant and make only those decisions that have not already been completed by machine-based modules.

Many human coding decisions (like correctly associating names, places, and dates with separate events reported in a single article) are facilitated by embedded NLP tools that populate the drop-down menus in the web interface used by coders to complete the event codings. For example, some drop-down menus are loaded with named entities from the article being analyzed so that the coder can identify which of those entities were targets or perpetrators of civil unrest. At the time a document is presented to coders, each sentence is further classified using NLP algorithms for relevance to civil unrest events. This sentence-level classifier—known as the Event Annotation Tool (EAT)—color-codes relevant sentences containing information relevant to the SSP.[8] Because all machine-extracted NLP information from the pre-processing stage merely serves as inputs for coders, humans make all final judgments about the accuracy and relevance of machine-generated information. The coded events are stored in a data archive and are accessible for statistical analysis. Documents categorized by humans as irrelevant (i.e., false-positives) are placed in a discard bin (see Figure 1).

What distinguishes SPEED's system as "progressive" is how its outputs are used for refining the machine-based modules. Within a progressive supervised learning system, the role of human coders is to reduce their menial work so they can focus on more cognitively challenging tasks. This is achieved by generating feedback that teaches computers to better replicate human decisions. To illustrate this point consider that a coder's first task in examining a screened document is to confirm the classifier's judgment – a task that requires about two minutes to complete. Our initial classifier was highly accurate at detecting irrelevant news stories, but less accurate at identifying relevant stories. While between 97% and 99% of the discarded documents were later confirmed by humans to contain no event-related information, only 33% of the documents sent to human coders contained relevant information. The other two-thirds were "false-positives." To improve the classification algorithm, around 60,000 documents that had been humanly processed were used as second wave of test/train data. The revised classifier increased the "true positive" rate from 33% to 87% among articles classified as relevant, while maintaining a "true negative" rate of 96% among articles classified as irrelevant – an improvement that generated enormous efficiencies.[9]

3.2 *SPEED's Societal Stability Protocol (SSP)*

The Societal Stability Protocol (SSP) [10] is designed to leverage SPEED's global news archive for the study of civil strife. It is organized around three categories of civil strife:

political expression (speeches, demonstrations, symbolic actions, etc.), *politically motivated attacks* (riots, kidnappings, shootings, etc.), and *disruptive state acts* (declarations of martial law, arrests of dissidents, censorship, etc.).[11] Data collection within the SSP is organized into six sections: *who* (actor characteristics), *what* (event type and consequences), *how* (mode of expression, weapons used), *where* (location and geo-physical setting), *when* (event date), and *why* (event origins). The types of events covered by the SSP and the data its collects speaks to the research agendas of a wide range of disciplines, from political science and sociology to psychology and anthropology. SSP data is of particular interest to sociologists concerned with such topic as the stated demands, motivating grievances, and political orientations of social movements (e.g., Walder 2009); the varieties and effects of state efforts to repress them (e.g., Earl 2011); the political consequences of civil unrest (e.g., Amenta et al. 2010; McAdam and Su 2002); and the temporal and spatial dimensions of social conflict (e.g., Owens, Su, and Snow 2013; Wagner-Pacifici and Hall 2012). All can be studied at global scale using SPEED event data.

Because the SSP generates information on hundreds of variables, developing its analytic potential requires constructing smaller subsets of composite variables. Two subsets are particularly important here. The first summarizes event origins by categorizing them as reflecting anti-government sentiments, socio-economic discontents, socio-cultural animosities, political desires and beliefs, desire to retain power, eco-scarcities, and so on.[12] The second subset includes seven event intensity measures that have been calculated using SSP data.[13] Three deal with events initiated by non-state actors: small-gauge expression, mass expression and political violence. Three others pertain to events initiated by state actors: the abuse of ordinary state powers, the initiation of extraordinary state acts, and political violence. The last gauges the intensity of coups.[14]

3.3 *Quality Control*

Because SPEED uses humans working with a complex protocol, it is crucial to ensure that they are highly trained and that they operate proficiently. Thus, coders begin their tenure by participating in an extensive training and testing regimen that requires nearly 70 hours to complete. This regimen includes lectures, one-on-one training, and group training sessions. Training culminates in a series of tests that gauge the coder's ability to implement the protocol in accord with established norms and understandings; the tests gauge their capacity to identify events and to code them properly. Trainees must pass these "gatekeeper tests" before they are allowed to generate production data. Reliability testing continues after coders begin production coding: they are blindly fed a set of pre-coded "test" articles at established intervals to detect slippages in reliability.[15]

4. Limits of Existing Civil Strife Data and the Value-added of Hybrid Systems

The value-added of hybrid data collection efforts are illustrated in the next two sections by comparing SPEED's rich event data to sparse and episodic event data.

4.1 *Sparse Data*

Sparse event data mask meaningful, within-category variation in destabilizing events. Ignoring these differences is troubling because the intensity of strife events varies considerably. Capturing those differences is important to gauging the threat posed to regimes as well as the scale of regime responses to those threats. Simply put, a non-violent protest by five people poses different issues from one involving 15,000 people – even though sparse event data equates them. Events also vary in their origins, location and timing – all of which are crucial to understanding such things as the reasons for civil strife, spatial diffusion patterns, and the pace at which the discontent unfolds. To illustrate the importance of event-specific differences this section focuses on event intensity; later sections demonstrate the importance of temporal and spatial differences, as well as origins.

To examine within-category differences in destabilizing events we use SSP data from a global random sample of *New York Times* stories that appeared between 1946 and 2005;[16] it contains records for over 70,000 codings.[17] Table 2 provides data on two types of political expression events: small-gauge events (e.g., provocative speeches/pamphlets, symbolic burnings) and mass expression events (e.g., demonstrations, strikes). The first column in Table 2 examines small-gauge expressions. An important intensity indicator for even small-gauge expression events is the number of participants. The mode and median for small-gauge events is one participant, but 10% involved more than 100 participants. Also important is the mode of expression. Verbal expressions and symbolic acts (e.g., sit-ins, self-immolations, pickets) each constituted 36%; another 25% are written expressions. Over 12% lasted more than a day and one-third were an integral part of a more complex sequence of actions; 10% involved some type of post-hoc reaction by a third-party (counter-demonstrations, arrests, attacks, etc.). The second column in Table

2 reports variation within mass events; almost 70% were demonstrations/ marches. Perhaps the most salient difference here is the number of participants. While the median is 2,450 and the mode is 1,250, the mean is almost 82,000; 10% of these events involved more than 95,000 people. Almost 30% lasted more than a day and more than two-fifths were part of a more complex sequence of actions.

TABLE 2

Table 3 provides descriptive data on political attacks initiated by private actors and state actors. Most attacks involve just a handful of initiators, but state-initiated attacks generally involve more.[18] About a third of political attacks are linked to a more complex sequence of actions. Much variance also exists in the weapons used. Nearly a third involved no weapon; moreover, non-state actors are more likely to use small arms and explosives while state actors are more likely to use military weapons. Personal attacks account for about 85% of the attacks, but injuries were reported in only about half. Egregious forms of violence such as mutilation or brutality occur in about 12% of attacks and are more likely to be perpetrated by non-state actors. Deaths occur in just over 50% of personal attacks; the median number killed is three.

TABLE 3

Most of the detail reported in Tables 2 and 3 would be lost given the current state of fully-automated systems. As the type of detail reported there is crucial to generating near-term advances in civil strife research, these data illustrate need for more sophisticated information extraction and the potential of hybrid approaches in providing it.

4.2 *Episodic Data*

Conventional human-centric approaches to episodic event data share two limitations (see Table 1). First, they are constructed using holistic judgments that are often derived from poorly-documented sources. In contrast, rich civil strife event data are systematically collected and aggregated from known populations. Second, the aggregation of discrete events into episodes often masks meaningful variation in conflict intensity over time and across space; rich event data can capture that variation. The next two sections illustrate the value-added of rich event data by examining the two key episodic strife variables: civil wars and periods of political instability.

4.2.1 *Civil Wars*

A number of prominent research projects have generated episodic data on the existence of a civil war, with the country-year as the unit of analysis: UCDP/PRIO (Gleditsch et al. 2002, www.prio.no); the Correlates of War Project (www.correlatesofwar.org); and a project directed by Fearon and Laitin (2003, www.stanford.edu/group/ethnic/publicdata). The attention accorded civil wars is understandable: they are the most devastating form of civil strife. But, for three reasons, an exclusive civil war focus is unlikely to advance our understanding of civil strife. First, measurement efforts have been hampered by a paucity of data and a lack of consensus on what constitutes a civil war. This has led to high levels of disagreement across measures. For example, during the 1945-1999 period a total of 1,272 country/years of civil war are identified by at least one of these three projects. However, all three agree on only 28% (357 country-years), which undermines confidence about inferences drawn from these datasets.

Even if scholars generated a definitive body of data on civil war battles and reached a consensus on what constitutes a civil war at the country-year level, such an operationalization would have limited utility as it would mask important patterns of temporal and spatial variation in strife. This can be illustrated using SSP data from a project using the SWB news archive to capture all relevant events from documents mentioning Guatemala, El Salvador, Nicaragua, Liberia, the Philippines, and Sierra Leone between 1979 and 2008 (see Rhodes et al. 2011). Figure 2 (a-d) aggregates, by month, the number of conflict-related deaths for four of these countries; the lines at the top of each graph mark the periods defined by the three civil war projects. Figure 2 suggests that much temporal variation exists in civil war conflicts. Some months show many casualties, but there are no reported deaths in about 60% of the months where there is unanimous agreement among the projects on the existence of a conflict. Most violence appears to stop well before the war is judged to have ended. While not shown in Figure 2, much spatial variance in the distribution of attacks also exists.[19]

FIGURE 2

Using more disaggregated death totals to gauge civil wars would be an improvement, but it would not address a third limitation: civil war variables capture only a small slice of civil strife perpetrated by a narrow set of actors – lethal violence caused by soldiers and insurgents. This can be illustrated by re-examining the SSP data used to produce Table 1 and 2. Political attacks constitute about 52% of these events. However, only 5% of all destabilizing events reported involved violent attacks between soldiers and insurgents. Moreover, if we consider only soldier and insurgent attacks where someone is killed (the core criterion for civil wars) the percentage drops to 2.6% of all destabilizing events.

4.2.2 *Episodes of Political Instability*

Research on political instability considers a broader range of event types than civil war research and the Political Instability Task Force (PITF) has done important work in this area (http://globalpolicy.gmu.edu/political-instability-task-force-home/). Recently, it introduced a model that predicted the outbreak of major periods of instability with a two-year lead-time (Goldstone et al. 2010). While the authors have a powerful and parsimonious model, they are vague about how they created their dependent variable, noting only that "We identified 'instability episodes' in part by identifying conflicts from existing databases (such as the Correlates of War) and in part by consulting with area experts" (Goldstone et al. 2010, 191).[20] Delineating instability episodes using methods that are rigorous and replicable is essential to advancing civil strife research and the opaqueness of PITF's holistic approach generates concerns about both.

PITF's definition of political instability includes civil wars, regime crises, and mass atrocities from 1955 to 2003. Thus, the 1946-2005 global random sample of strife events employed in Table 1 and 2 can be used to evaluate PITF's holistic approach. Our evaluative focus is on the specification of distinctive and cohesive episodes of instability;

it includes PITF's ability to *identify* distinctive sequences of instability and to *demarcate* them precisely. If PITF has not identified distinctive and cohesive episodes of instability then its dependent variable is poorly specified, which undermines confidence in its predictive model. Our concern with identification includes both false-positives (PITF-specified periods that are not cohesive and markedly unstable) and false-negatives (cohesive periods of marked instability that were not identified). With respect to demarcation our concern is with PITF's ability to accurately specify episodic "bookends" (i.e., start and end-points). If they cannot, then their claim of predicting the outbreak of instability with a two-year lead-time is unpersuasive.

To examine the PITF approach we aggregated the seven intensity measures introduced earlier – which capture everything from the intensity of political expression and political violence to state repression and coups – to the country-month level and merged them with PITF data. Then we reduced the seven measures to a weighted composite intensity variable. A score of '0' on this intensity measure indicates a month with no reported unrest; increasing values reflect higher levels of instability.[21] Figure 3 graphs data for six countries that illustrate how our evaluation of PITF analysis was conducted. In these graphs, the X-axis plots country-months from 1955 through 2004, PITF episodes are shaded in gray, and our composite intensity measure is plotted on the Y-axis.

Figure 3a, which reports data on Djibouti, illustrates an important type of false positive: episodes that do not appear to be distinctively unstable. While the demarcated period for Djibouti is a PITF episode, and is accorded the same value as all other episodes, its levels of instability are markedly lower than the others in Figure 3. Indeed, we find no recorded events during the PITF timeframe even though we find some major strife in Djibouti at other points in the postwar era. When the average intensity score of a PITF episode is less than the average intensity score for country-months falling outside PITF episodes, as is the case in Djibouti, we define the episode as a false positive. Fifty-five of the 145 PITF episodes (38%) fail to exceed this minimal intensity threshold and are excluded from the following analyses.[22] Figure 3b depicts intensity data for Indonesia and illustrates a second type of false-positive. While PITF data indicate that Indonesia was experiencing a good deal of instability during the PITF-defined timeframe, SPEED data show intermittent instability with long interludes of calm. This suggests that several distinct episodes may be merged into one. We defined PITF episodes with interludes of calm that exceed two years – which is longer than 90% of the interludes in the PITF episodes – to be false positives. Thirty-two of the 145 episodes (22%) were affected by at least one interlude exceeding this interlude threshold; eight of these had more than one interlude.

FIGURE 3

Figure 3c and 3d provide illustrations of false-negatives: cohesive sequences with average intensity levels that match those in PITF episodes yet are not captured in a PITF episode. Figure 3c shows that PITF captures only one of several important periods of instability in Egypt. Figure 3d shows that while Bolivia does not have a single PITF episode, it has several periods of strife that far exceed the average level for PITF episodes (m = 2.45). Using criteria derived from the average value of SPEED’s intensity variables within PITF episodes and applying them to country-months not included in a PITF episode, we found 979 additional episodes of civil strife.[23] While PITF captured the most salient episodes of political instability in the postwar era, the additional episodes uncovered using SPEED data compare favorably with the PITF episodes in terms of intensity levels. Consider, for example, that the median intensity score for the 979 false-negatives is slightly higher than that for comparable episodes derived from the PITF episodes: 3.3 (m=8.6) vs. 3 (m= 6.45). In contrast, the duration of validated PITF episodes is somewhat longer than for the false-negatives: 11.5 months (mean=32.5) for true-positive PITF episodes vs. 1 month (mean=8.7) for the false-negative episodes revealed in SPEED data. Finally, 4.8% of the country-months included in the false-negatives involved some type of coup activity, which is somewhat higher than the 4.3% found in validated PITF episodes.

The last component of our analysis pertains to the accuracy with which the PITF approach demarcates episodic start- and end-points. Start-points are particularly important because they affect PITF's ability to predict eruptions of instability. Figures 3e and 3f illustrate the demarcation analysis. Figure 3e shows that the SPEED data for Lebanon spill beyond PITF's temporal boundaries; Figure 3f shows that the instability in Sierra Leone begins well after the PITF episode starts and ends well before PITF's endpoint. To quantify the accuracy of PITF's bookends we used a six-month criterion, which is 25% of the lead-time that PITF employs in its analysis. We found that the start-points were mis-specified (i.e., off by at least six months) in 18 of the 90 validated PITF episodes; end-points were mis-specified in 40 validated episodes.

In sum, our analysis suggests that 55 of the 145 PITF episodes involve sequences of country-months that are not distinctively different from the country-months that fall outside PITF episodes. Of the remaining 90 PITF episodes, another 32 included interludes of calm of at least two years and did not constitute cohesive sequences of on-going conflict. Fifty-eight episodes had mis-specified bookends. After eliminating overlaps among the different types of error, SPEED data suggests that 107 of the 145 original PITF episodes had some type of serious measurement error. Even more troubling is what PITF failed to uncover: 979 episodes that had levels of strife comparable to the validated PITF episodes. No dependent variable in the social sciences is free of measurement error, but the level of noise in the PITF measure underscores the value-added of rich event data in civil strife research.

5.0 New Frontiers in Civil Strife Research

The value of SPEED's rich strife data goes beyond its methodological advantages. The strategic use of these data can advance the frontiers of civil strife research and yield fresh substantive insights. To illustrate this point we join two sets of measures introduced earlier, our intensity and origins composites. Our efforts to glean the origins of individual events from news reports suggest that most destabilizing events are rooted in common grievances that vary in prominence across event type, space and time. Comparing the intensity of events associated with different types of grievances enhances our understanding of the changing nature of contentious politics in the postwar era by generating more refined global insights into the type of grievances that are driving strife, how those drivers have changed over time, and the changes in how those discontents are manifested. This can be illustrated by examining the two most disruptive forms of citizen-initiated strife: mass expression and political violence.

To depict the relative importance of the different origins, and how they vary over time, Figure 4 displays a set of Lowess regression lines that track the global prominence of different grievances from 1946 to 2005. Because the range in the intensity of these two types of strife events is much different, we use different scales to depict them; moreover, for succinctness we graph only the top four drivers of unrest: anti-government sentiments, socio-cultural animosities, socio-economic discontents and the desire for enhanced political rights. Figure 4 (a) shows that the most important drivers of mass expression are anti-government sentiments and socio-economic discontents, but that their relative prominence varies over time. In the late 1940's socio-economic concerns were the most important factor. But the role of socio-economic discontents declines precipitously over time and by the end of the period it is the weakest driver.[24] In contrast, anti-government sentiments grow as a driver of mass expression and are the most important driver after the mid-1950s. Socio-cultural animosities also play an increasingly important role in mass expressions from the mid-1950s through the late 1990s. The role of the desire for political rights is fairly stable throughout the timeframe but it ebbs and flows.

FIGURE 4

Figure 4 (b) shows that the use of violence to express discontent evidences a somewhat different pattern. The principal drivers of political violence are socio-cultural animosities and anti-government sentiments. Both evidence a relative decline until the early 1960s. But after that point socio-cultural animosities become much more potent until the mid-1980s, when they begin to recede. In contrast, anti-government sentiments manifest a fairly stable pattern until the mid-1980s when they evidence a decline. Both drivers demonstrate a slight upturn at the turn of the century. The desire for enhanced political rights is an important and largely stable factor throughout the period but it, too, begins to decline in the mid-1980s until it upticks around 2000. In contrast socio-economic discontents, which were comparable in potency to the desire for enhanced political rights at the start of the time frame, decline steadily over time.

6.0 Summary

The revolution in information technologies – both by generating "Big Data" and tools to transform those data into knowledge – presents enormous opportunities for social scientists. The vast increases in computational capacities, combined with the adoption of data science techniques, can create exciting new research frontiers that will transform the social sciences in the same way that the molecular revolution transformed biology. Perhaps the only social science parallel to these contemporary developments is the widespread diffusion of telephones and the refinement of sampling techniques and survey methods after WWII. However, despite the enormous promise of the information revolution for social science, the trajectory forward is not likely to be linear or steep. The need to derive meaning from complex language patterns and the current state of data science techniques for analyzing unstructured data suggest that social scientists will continue to balance the relative advantages of machine-based and human-centric approaches into the foreseeable future. The central assertion of this article is that, until machine-based approaches can more accurately emulate human-centric approaches, researchers should consider the use of hybrid approaches that strategically integrate the benefits of both.

The rationale for this assertion is that, while wholly human-centric approaches will never realize the potential of the information revolution, the premature embrace of fully automated approaches when studying complex social phenomena will sacrifice validity and nuance to achieve scale. This paper introduced one hybrid approach (the SPEED project) and demonstrated its value-added in a complex domain (civil strife) that is of interest to an array of social scientists. Notwithstanding the results reported here, there are several reasons social scientists might justifiably continue employing human-centric analyses; three are particularly important.

First, the standard machine-based techniques for processing textual data work only when applied at scale to extremely large textual corpora. Many social scientists work with relatively small corpora (on the order of tens or hundreds of texts), and few computational approaches work with such small numbers. Second, computational approaches require data pipelines so complex and programming expertise so specialized that text-mining systems essentially function as black boxes to the non-expert. Validating or bias-checking the process by which these workflows transform text into data is usually infeasible, even with publicly released code. In contrast, traditional content analysis methods involve the disclosure of codebooks and well-developed standards for assessing validity and reliability, making it straightforward for non-specialists to evaluate data quality. Third, text-mining systems are so costly to build that even if the software components are open-source and distributed without cost, they are hardly “free.” Lexicoder (www.lexicoder.com), RapidMiner (www.rapidminer.com), and R (www.r-project.org) are three of many such components distributed free of charge, but the opportunity costs are steep for deploying them effectively.

The SPEED project illustrates the scale of these opportunity costs. The Cline Center began assembling its news archive and developing SPEED’s workflow system in 2006, but lacked an operational cyberinfrastructure until 2009. Seven years and well over a million dollars later, the Cline Center released its first SPEED data set. Opportunity costs and resource demands on this scale are formidable but they need not deter serious scholars as there is no need to duplicate the Cline Center’s foundational efforts. While agreements with commercial vendors and intellectual property rights prohibit the Center from distributing its news archive, efforts are being made to provide non-consumptive public access to the Center’s holdings. This access will allow researchers to evaluate the utility of the Center’s digital archive for their needs and construct a research design to realize those needs. Based on that design, researchers can utilize the Center’s various sub-centers of expertise (document classification, training, coding, etc.) to implement it.

The benefits of extending the use of SPEED’s hybrid system to a broader range of scholars, even at this early stage of data science, are substantial. In our view the most immediate research benefits lie within three areas: document classification; text annotation; and the clustering of documents about the same event from different sources. Classifying 5.9 million *New York Times* articles on the basis of civil unrest content would have taken a single human analyst working 24 hours a day and 365 days a year over two decades to complete. Once SPEED's classifier model was fine-tuned, this task was completed in a matter of hours. Using NLP solutions to pre-annotate relevant text so that humans can quickly scan for relevant content is vital to analyzing large numbers of classified documents efficiently. If SPEED's EAT module eventually reduces processing time by just two minutes an article, the time required to process all *New York Times* civil strife articles would be reduced by almost 35,000 person-hours. Identifying articles on different news websites that describe the same event using different terms is the key to capitalizing upon the diversity of news websites and overcoming the biases introduced by relying on a small number of outlets. To utilize diverse sources without clustering reports of similar events will generate duplicate codings and undermine the validity of the data. SPEED has yet to finalize its approach to this problem, but finding a solution is at the top of its current development agenda.

As data science methods mature their contributions will be even more profound and far-reaching, further exacerbating the quandary faced by social scientist dealing with the opportunity costs of integrating automated components into content analysis projects. Based on our experience with SPEED we believe that the most propitious path forward is to create collaborations between social scientists and data scientists. It is through such collaborations that social scientists will be able to capitalize on data science techniques while retaining the nuance needed for studying social complex phenomena. These collaborative efforts exploit a mutually beneficial division of labor across academic disciplines and are a highly efficient way of employing automated techniques to generate important social science payoffs. Social scientists should be willing collaborators in these efforts as they have much to contribute to, and much to gain from, such joint enterprises.

[1] Other automated efforts, such as DARPA’s ICEWS (O'Brien 2010), are based on technologies developed by these other projects.

[2] A perusal of its data archive suggests that it draws from hundreds of sources, including such diverse outlets as the New York Times, Reuters News, BBC Monitoring Service, Africa Research Notes, and African Contemporary Reports.

[3] This Heretrix crawler (v.3.1.1) is an open-source platform that retrieves news articles published from a set of monitored RSS feeds. Before moving to the Heretrix platform, SPEED web data were collected using a scratch-built RSS crawler. Data retrieved through the Heretrix system is sent through several cleanup steps before storage. We use Google’s open-source lang-detect software (http://code.google.com/p/language-detection) to exclude any crawled news stories in languages other than English. We also attempt to remove all html tags as well as header, footer, and copyright data from the crawled content, so that only the raw text of the news article is stored. We then perform near duplicate detection (NDD) to eliminate duplicate copies of the same article that might have been retrieved from different RSS feeds. NDD is performed using a shingle approach that creates a hash for 50 characters created as a moving window across the text, saving 100 shingles for each document. The shingles for each document are saved for five days. Each incoming document is compared to every other document using the shingles that are currently saved and eliminated if the matching score is greater than 80%. Additional details on SPEED’s global news archives can be found at: http://www.clinecenter.illinois.edu/research/documents/SPEEDInformationBase.pdf

[4] Useful overviews of this field can be found in Murphy (2012); Hastie, Tibshirani, and Friedman

(2009); and Abu-Mostafa, Magdon-Ismail, and Lin (2012).

[5] Before settling on a Naïve Bayes algorithm, testing was also conducted using Support Vector Machine, Decision Tree, and Neural Network classifiers. Naïve Bayes using a Term Frequency model outperformed the rest. More information on this classification system (labeled the **BIN** system) is provided at the following address: http://www.clinecenter.illinois.edu/research/publications/SPEED-BIN.pdf. SPEED currently uses RapidMiner 5.3.008 open-source software to implement the classifier system.

[6] Human re-analysis of the discarded news stories confirmed that nearly all of them were correctly classified: between 1% and 3% of the discards were found by human analysts to contain event-related information, which was considered an acceptable false-negative rate.

[7] The protocols make extensive use of drop-down lists, response-activated questions, and branching commands that "hide" irrelevant questions. Moreover, NLP-based techniques automatically capture proper names, facilitate the identification of dates, and aid in the integration of geo-spatial data.

[8] EAT was incorporated into SPEED's workflow in 2014 after a five-year collaboration with the Cognitive

Computation Group (http://cogcomp.cs.illinois.edu/) at the University of Illinois. The collaboration with CCG continues as efforts are currently being made to incorporate feedback loops to improve EAT's algorithms. More information on EAT is available in the online supplemental material.

[9] The revised classifier used a Naïve Bayes algorithm with a Term Frequency model, as did the initial classifer. We used 10-fold cross validation with the final model using all 10 models. The Term Frequency model uses a set of 5,673 words for the modeling. The average accuracy of the model was 84.03%. The false negative and positive are 3.258% and 12.707% respectively which is a significant improvement by reducing both errors by about 50% over the initial models built.

[10] We have worked on SPEED protocols for a variety of applications (security of property rights, integrity of elections, expressive freedoms). However, because of the increasing importance of civil strife most of our developmental work has focused on the SSP. The architecture of the SSP is detailed in a white paper that can be found at:

http://www.clinecenter.illinois.edu/research/documents/AnOverviewoftheSSP.pdf

[11] The classifier used for the SSP generates statistical probabilities that a news report contains information on at least one event that falls within the SSP's three-category event ontology. A document detailing the operational definitions of each type of event in this ontology can be accessed at the following address: http://www.clinecenter.illinois.edu/research/publications/SPEED-Definitions_of_Destabilizing_Events.pdf.

[12] The derivation of the origins composites is reported at: http://www.clinecenter.illinois.edu/research/publications/SPEED-Origins_of_Destabilizing_Events.pdf.

[13] A detailed discussion of how these intensity measures were derived can be found at: http://www.clinecenter.illinois.edu/research/publications/SPEED-Gauging_the_Intensity_of_Civil_Unrest.pdf.

[14] The coup data comes from the Coup d'état project: http://www.clinecenter.illinois.edu/research/documents/Coup_Project.pdf, for reasons outlined in note 4 of the online supplemental material.

[15] A fuller discussion of SPEED's training and testing procedures, including the results of the reliability tests can be found at: http://www.clinecenter.illinois.edu/research/publications/SPEED-Reliability.pdf.

[16] This sample was generated by randomly sampling every *New York Times* article that our automated classifier indicated had content relevant to a civil strife event.

[17] Moving from unstructured media data to probabilistic estimates of various properties and relationships requires addressing a number of well-known challenges in a methodologically responsible manner. Advances in information technology have provided researchers with the potential to address some of these challenges and the SPEED project has made a concerted effort to exploit this potential; its efforts are detailed at: http://www.clinecenter.illinois.edu/research/publications/Media%20Data%20and%20Social%20Science%20Research.pdf. Particularly relevant for the cross-national analyses included in this section are SPEED's

efforts to address the country biases embedded in the *New York Times*. An examination of the random sample of events drawn from the Times revealed a considerable bias in coverage toward rich, Western democracies and global competitors to the U.S. (Soviet Union, China, etc.).

To address this bias we employed a weighting scheme that was implemented at the country-month level. We used NLP techniques to identify, for each month between January 1, 1946 and December 31, 2005: (1) the number of New York Times articles in which each country (or a city or province in the country) was mentioned; and (2) the total number of news articles published. The proportion of total articles in a month that included a reference to a country was then calculated for each country-month. That proportion formed the denominator for the weighting scheme. Next, using linear interpolation, we calculated populations for each country-month and expressed them as a proportion of the global population. That proportion was the numerator for the weighting scheme.

The effect of this scheme is to reduce the impact of country-based media bias on our analysis. To illustrate, consider a country that constituted only 1% of the global population in a given month, yet was mentioned in 4% of the NYT articles in that month. Its weight would be .25. In contrast, a country that constituted 4% of the world's population, but was mentioned in only 1% of the New York Times articles, would have a weight of 4. This weighting scheme was used in all of the cross-national analyses reported in this section. A more complete description of the weighting scheme and its justification is provided in the white paper at the above cited website.

A major assumption underlying the weighting scheme used here is that at least some of the destabilizing events that unfold in an enduring episode of strife are reported in the media. If none are reported the use of the weights developed here will not mitigate country-based media bias. This assumption may not hold with respect to some types of events (community meetings, small, non-violent demonstrations, isolated incidents of minor violence, etc.) but we feel comfortable in using it in the analysis of enduring episodes of strife for two reasons. The first is that, by definition, enduring episodes of strife involve a *series* of salient events, at least some of which are violent. This greatly enhances the likelihood of media coverage for at least some of the period covered by the strife episode. The second is the widespread use of recapitulation passages by print media. Recapitulation passages are bodies of texts that summarize past happenings that are relevant to a story (for more information on these passages see: http://www.clinecenter.illinois.edu/publications/SPEED-Transforming_Textual_Information.pdf: p. 7). The use of this reporting technique allows the news media to play "catch-up" that enhances coverage in less salient countries; recapitulation codings account for nearly 12% of the codings used in the database used to identify enduring episodes of strife. The fact that we identify multiple episodes of strife in such obscure countries as Comoros, Madagascar, Lesotho, Namibia and the Central African Republic suggests that the weighting scheme devised here works effectively for its intended purpose.

[18] Although not shown in Table 3, the vast majority of state attacks (87%) were initiated by soldiers or police officers, while only half of non-state attacks are initiated by insurgent groups.

[19] A Google Earth display showing the spatial distribution of SSP events in these countries can be found at: http://www.clinecenter.illinois.edu/research/speed-data.html.

[20] While the PITF website includes thumbnail descriptions of their episodes, they do not include concrete criteria for inclusion. The descriptions are at http://www.systemicpeace.org/inscr/PITF%20Consolidated%20Case%20List2011.pdf.

[21] For interested readers, a detailed explanation of the reformatting of the intensity measures and the derivation of the composite intensity measure is provided at: http://www.clinecenter.illinois.edu/research/documents/SSP_Episodes_Demarcation.pdf.

[22] The mean intensity score for the other PITF episodes is 2.68 and the median is .63. In contrast, the mean value for the fifty-five episodes identified here is just .03 (median=.01), a markedly lower level of unrest intensity compared to the other episodes.

[23] Details on the procedures used to identify these episodes can be found at: http://www.clinecenter.illinois.edu/research/documents/SSP_Episodes_Demarcation.pdf

[24] A different picture would likely emerge if the time line was extended to capture the effects of the 2008 global recession.

References

Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. 2012. *Learning from data: A short course*: AMLBook.com.

Althaus, Scott L., Nathaniel Swigger, Svitlana Chernykh, David Hendry, Sergio Wals, and Christopher Tiwald. 2011. Assumed transmission in political science: A call for bringing description back in. *Journal of Politics* 73 (4):1065-1080.

Amenta, Edwin, Neal Caren, Elizabeth Chiarello, and Yang Su. 2010. The Political Consequences of Social Movements. *Annual Review of Sociology* 36 (1):287-307.

Armstrong, J. Scott. 1967. Derivation of Theory by Means of Factor Analysis or Tom Swift and His Electric Factor Analysis Machine. *The American Statistician* 21 (5):17-21.

Berelson, Bernard. 1952. *Content analysis in communication research*. New York: Hafner.

Bernauer, Thomas, and Nils Petter Gleditsch. 2012. New Event Data in Conflict Research. *International Interactions* 38 (4):375-381.

Bond, Doug, Joe Bond, Churl Oh, J. C. Jenkins, and Charles Lewis Taylor. 2003. Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development. *Journal of Peace Research* 40 (6):733-745.

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.

Conway, Mike. 2006. The subjective precision of computers: a methodological comparison with human coding in content analysis. *Journalism and Mass Communication Quarterly* 83 (1):186-200.

Danzger, M. Herbert. 1975. Validating conflict data. *American Sociological Review* 40 (5):570-584.

Earl, Jennifer. 2011. Political Repression: Iron Fists, Velvet Gloves, and Diffuse Control. *Annual Review of Sociology* 37 (1):261-284.

Eck, Kristine. 2012. In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. *Cooperation and Conflict* 47 (1):124-141.

Evans, Michael, Wayne McIntosh, Jimmy Lin, and Cynthia Cates. 2007. Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research. *Journal of Empirical Legal Studies* 4 (4):1007-1039.

Fearon, James D., and David D. Laitin. 2003. Ethnicity, Insurgency and Civil War. *American Political Science Review* 97:75-90.

Franzosi, Roberto. 1987. The press as a source of socio-historical data: Issues in the methodology of data collection from newspapers. *Historical Methods* 20 (1):5-16.

Repeated Author. 2004. *From words to numbers: Narrative, data, and social science*. New York: Cambridge University Press.

Franzosi, Roberto, Gianluca De Fazio, and Stefania Vicari. 2012. Ways of Measuring Agency: An Application of Quantitative Narrative Analysis to Lynchings in Georgia (1875-1930). *Sociological Methodology* 42:1-42.

Gerner, Deborah J., Philip A. Schrodt, Rajaa Abu-Jabr, and Omur Yilmaz. 2002. Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Intearctions. In *43rd Annual Convention of the International*

Studies Association. New Orleans, LA.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Havard Strand. 2002. Armed Conflict 1946-2001: A New Dataset. *Journal of Peace Research* 39:615-637.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustick, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. 2010. A Global Model for Forecasting Political Instability. *American Journal of Political Science* 54 (1):190-208.

Grimmer, Justin, and Gary King. 2011. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*.

Grimmer, Justin, and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21 (3):267-297.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. New York: Springer.

Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2008. Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics* 4 (4):31 - 46.

Holsti, R. 1964. An adaptation of the "General Inquirer" for the systematic analysis of political documents. *Behavioral science* 9 (4):382-8.

Jackman, Robert W., and William A. Boyd. 1979. Multiple sources in the collection of data on political conflict. *American Journal of Political Science* 23 (2):434-458.

Kahl, Colin H. 2006. *States, Scarcity and Civil Strife in the Developing World*. Princeton and Oxford: Princeton University Press.

King, Gary, and Will Lowe. 2003. An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization* 57 (3):617-642.

Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology*. 2nd ed. Thousand Oaks, CA: Sage.

Lacina, Bethany, and Nils P. Gleditsch. 2005. Monitoring trends in global combat: A new dataset of battle deaths. *European Journal of Population* 21 (2):145-166.

Liu, Bing. 2011. *Web data mining: Exploring hyperlinks, contents, and usage data*. 2nd ed. Berlin: Springer-Verlag.

Lohr, Steve. 2013. Algorithms Get a Human Hand in Steering Web. *New York Times*.

McAdam, Doug, and Yang Su. 2002. The War at Home: Antiwar Protests and Congressional Voting, 1965 to 1973. *American Sociological Review* 67 (5):696-721.

Monroe, Burt L., and Philip A. Schrodt. 2008. Introduction to the Special Issue: The Statistical Analysis of Political Text. *Political Analysis* 16 (4):351-355.

Murphy, Kevin P. 2012. *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

National Consortium for the Study of Terrorism and Responses to Terrorism. 2012. Global Terrorism Database [Data file]. Retrieved from http://www.start.umd.edu/gtd.

Neuendorf, Kimberly A. 2002. *The content analysis guidebook*. Thousand Oaks, CA: Sage.

O'Brien, Sean P. 2010. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review* 12 (1):87-104.

Owens, Peter B., Yang Su, and David A. Snow. 2013. Social Scientific Inquiry Into Genocide and Mass Killing: From Unitary Outcome to Complex Processes. *Annual Review of*

Sociology 39 (1):69-84.

Potter, W. James, and Deborah Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research* 27:258-284.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science* 54 (1):209-228.

Raleigh, Clionadh. 2012. Violence Against Citizens: A Disaggregated Analysis. *International Interactions* 38 (4):462-481.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing ACLED: An Armed Conflict Location and Event Dataset. *Journal of Peace Research* 47 (5):651-660.

Rhodes, A., A. Waleij, W. Goran, A. Singh, and P. Nardulli. 2011. Proactive Peacebuilding with Natural Resource Assets. U.S. Army Corps of Engineers; Center for the Advancement of Sustainability Innovations.

Salehyan, Idean, Cullen S. Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Sull, and Jennifer Williams. 2012. Social Conflict in Africa: A New Database. *International Interactions* 38 (4):503-511.

Savage, Charlie. 2013. N.S.A. said to search content of messages to and from U.S. *New York Times*, August 8, A1.

Schrodt, Philip A. 2012. Precedents, Progress, and Prospects in Political Event Data. *International Interactions* 38 (4):546-569.

Soroka, Stuart N. 2012. The Gatekeeping Function: Distributions of Information in Media and the Real World. *The Journal of Politics* 74 (02):514-528.

Stone, Philip J. 1962. *The general inquirer : a computer system for content analysis and retrieval based on the sentence as a unit of information*. Harvard: Laboratory of Social Relations, Harvard University.

Sudhahar, Saatviga, Gianluca De Fazio, Roberto Franzosi, and Nello Cristianini. 2013. Network analysis of narrative content in large corpora. *Natural Language Engineering* 20:1-32.

Themner, Lotta, and Peter Wallensteen. 2011. Armed Conflict, 1946-2010. *Journal of Peace Research* 48 (4):525-536.

Themnér, Lotta, and Peter Wallensteen. 2011. Armed conflict, 1946–2010. *Journal of Peace Research* 48 (4):525-536.

Thompson, Bruce. 1995. Stepwise Regression and Stepwise Discriminant Analysis Need Not Apply here: A Guidelines Editorial. *Educational and Psychological Measurement* 55 (4):525-534.

Urdal, Henrik, and Kristian Hoelscher. 2012. Explaining Urban Social Disorder and Violence: An Empirical Study of Event Data from Asian and Sub-Saharan African Cities. *International Interactions* 38 (4):512-528.

van Atteveldt, Wouter, Jan Kleinnijenhuis, Nel Ruigrok, and Stefan Schlobach. 2008. Good News or Bad News? Conducting Sentiment Analysis on Dutch Text to Distinguish Between Positive and Negative Relations. *Journal of Information Technology & Politics* 5 (1):73 - 94.

Wagner-Pacifici, Robin, and Meredith Hall. 2012. Resolution of Social Conflict. *Annual Review of Sociology* 38 (1):181-199.

Walder, Andrew G. 2009. Political Sociology and Social Movements. *Annual Review of Sociology* 35 (1):393-412.

Witten, Ian H., Eibe Frank, and Mark A. Hall. 2011. *Data mining: Practical machine learning tools and techniques*. 3rd ed. Amsterdam: Morgan Kaufmann Publishers.

Woolley, John T. 2000. Using media-based data in studies of politics. *American Journal of Political Science* 44 (1):156-173.

Young, Lori, and Stuart Soroka. 2012. Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication* 29 (2):205-231.

Figure 1
A Progressive, Supervised-learning System

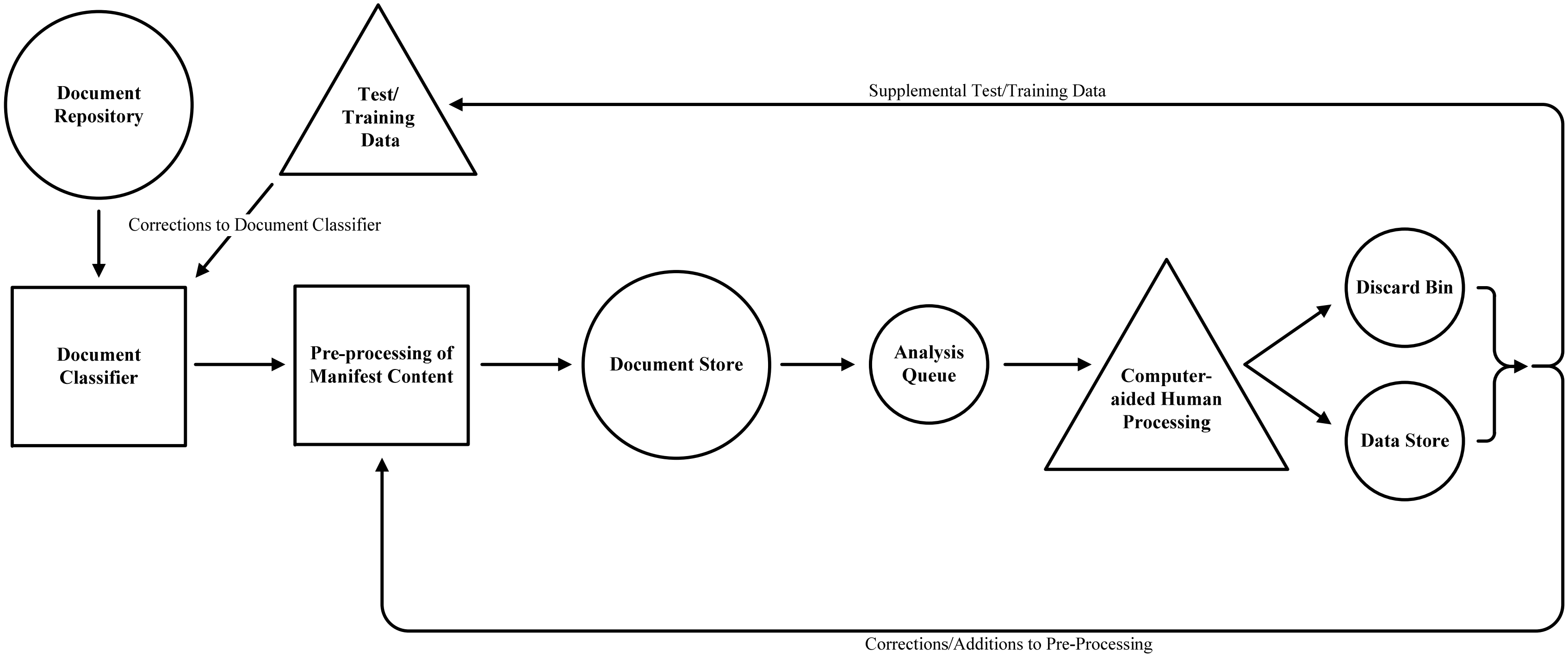
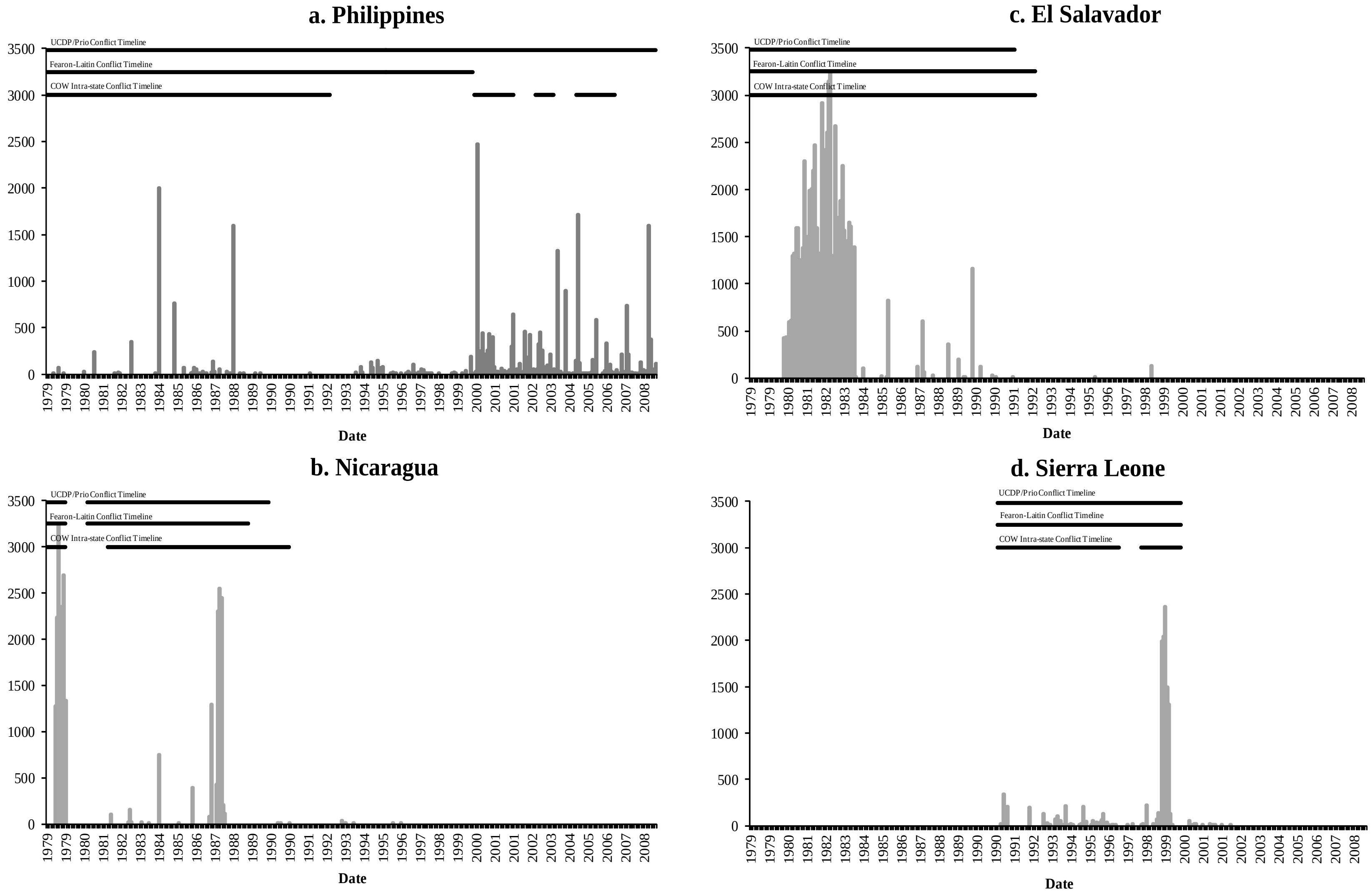

Document Repository
Test/ Training Data
Supplemental Test/Training Data
Corrections to Document Classifier
Document Classifier
Pre-processing of Manifest Content
Document Store
Analysis Queue
Computer-aided Human Processing
Discard Bin
Data Store
Corrections/Additions to Pre-Processing

Figure 2
Monthly Death Totals for Soldiers and Insurgents



a. Philippines

b. Nicaragua

c. El Salavador

d. Sierra Leone

Figure 3
Overlay of Selected PITF Episodes and SSP Monthly Aggregates

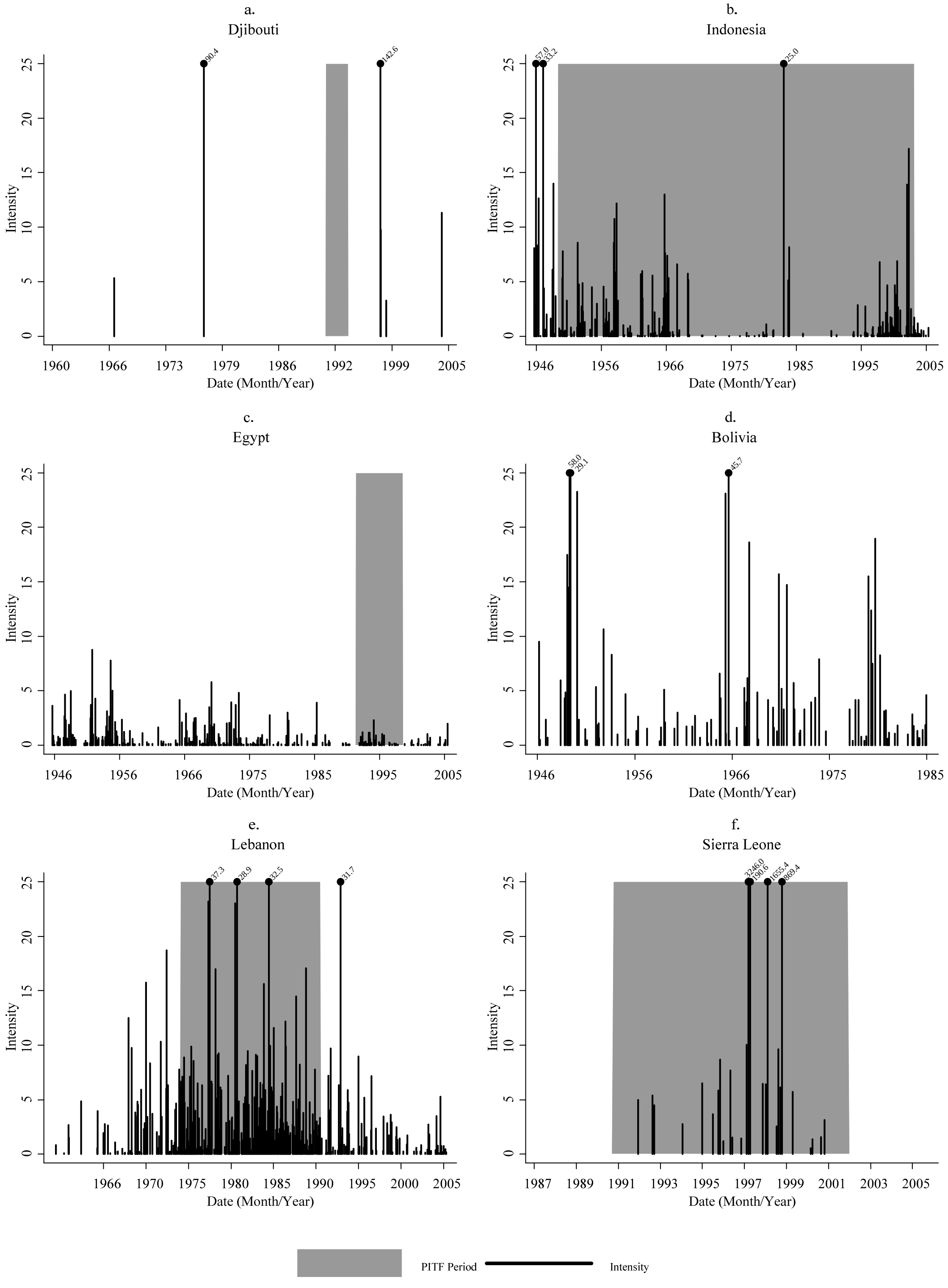
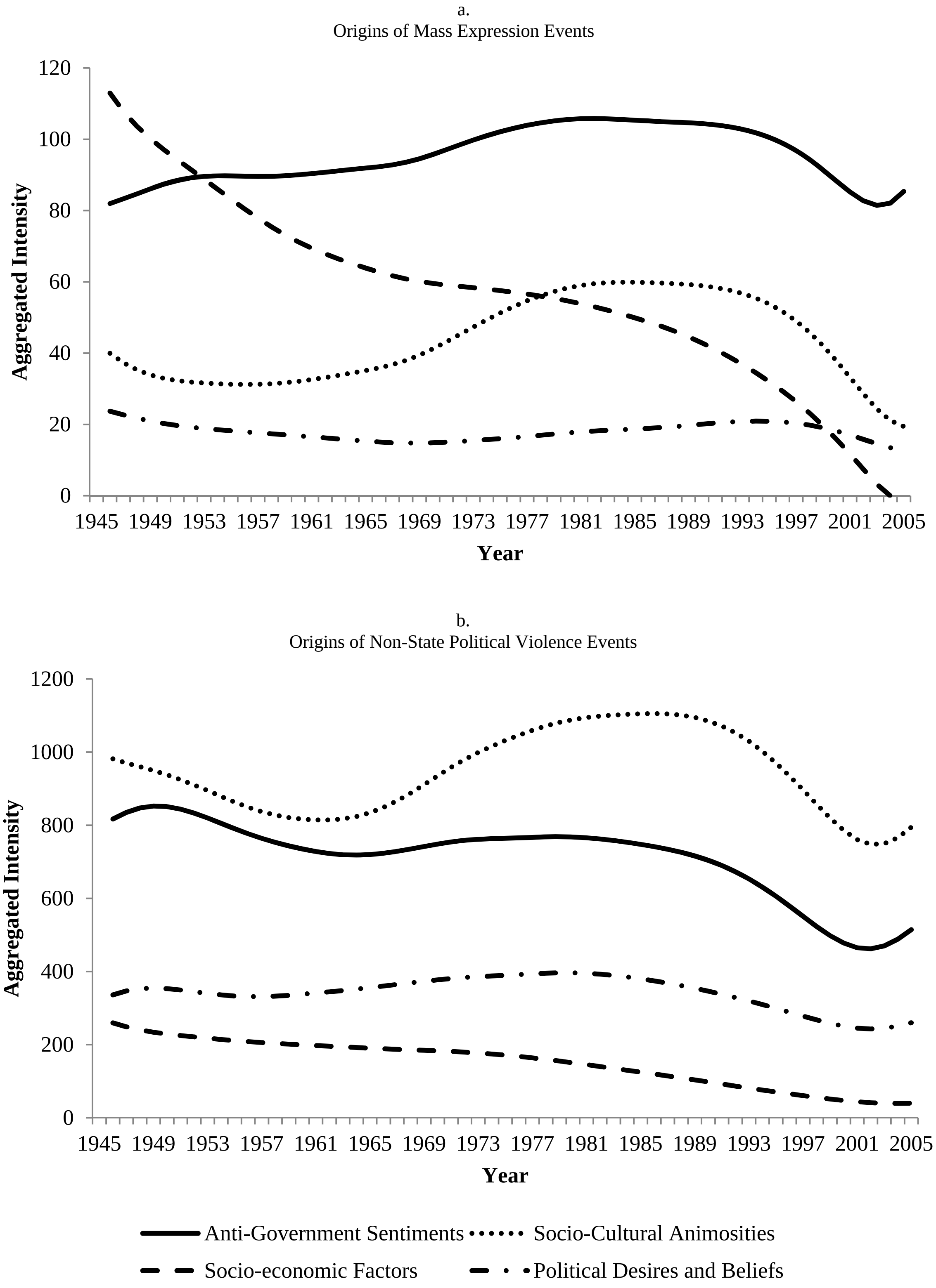


Figure 4
Origins of Mass Expressions and Non-state Political Violence



a.
Origins of Mass Expression Events

b.
Origins of Non-State Political Violence Events

Table 1
Types of Civil Unrest Event Data

	Episodic Data	**Discrete Data**	
		Sparse	*Rich*
Typical Spatial Resolution	Country	Country	City
Typical Temporal Resolution	Year	Day	Day
Level of Detail	Whether a state of civil war/political instability exists or not	Whether a type of event occurred or not	Event intensity, origins, and outcome; number and identity of perpetrators/targets; linkage to previous events
Examples	UCDP/PRIO, COW	CAMEO, IDEA	ACLED, GTD
Largely Automated	No	Yes	No
Advantage	Synthesizes discrete events into conflict periods even when event-specific data are lacking	Scale and speed of data collection using automated methods	Contextual data adds precision and depth to analyses
Disadvantage	Treats all conflicts as equivalent regardless of scale or intensity; subjective judgments used to identify and demarcate episodes	Treats all events of a given type as equivalent regardless of scale or intensity	Human involvement in data collection process limits scale and speed of data collection

Table 2
Within-category Variance in Expression Events

	Small-scale Expression Events	**Mass Expression Events**
Number of Participants/Initiators		
Mean	555	81,982
Median	2	2,450
Mode (proportion)	1 (.48)	1,250 (.17)
Mode of Expression		
Verbal	.36	.
Written	.25	.
Symbolic Action	.36	.
Demonstration	.	.69
Strike	.	.31
Intensity Indicators		
Part of a Sequence of Events	.34	.42
Elicited a Posthoc Reaction	.10	.06
Lasted More than a Single Day	.12	.29
N =	7,409	8,663

Table 3
Within-category Variance in Political Attacks

	Non-State Attacks	State Attacks
Number of Initiators		
Mean	634	1022
Median	4-5	4-5
Mode (proportion)	5 (.28)	5 (.56)
Relationship to Other Events		
Involved a Linked Event?	.41	.30
Elicited a Post-hoc Reaction?	.04	.14
Weapon Type		
None	.30	.33
Crude	.11	.06
Small Arms	.25	.20
Explosives	.31	.03
Military Grade	.03	.38
Type of Violence		
Attack Against Person	.81	.88
Personal Injury	.51	.52
Egregious Violence	.14	.08
Lethal Injuries		
Proportion Involving a Death	.40	.49
Mean deaths (for lethal events)	1859	1780
Median deaths (for lethal events)	2	5
Modal deaths (for lethal events)	1 (.31)	1 (.25)
N =	18,860	7,340