GDELT: Global Data on Events, Location and Tone, 1979-2012. *

Kalev Leetaru
kalev.leetaru5@gmail.com

Philip A. Schrodt
schrodt@psu.edu

Version 1.0 : March 29, 2013

*Paper presented at the International Studies Association meetings, San Francisco, April 2013. The authors would like to specifically acknowledge the following organizations in making this research possible: BBC Monitoring, Reed Elsevier's LexisNexis Group, and Google and Google News. We are also indebted to [Dr.] Jay Yonamine for the visualizations of the Syria and Afghanistan included in this paper, and extensive experimentation with GDELT in his recently-defended dissertation. Schrodt's contributions to the project were funded in part by National Science Foundation grant SES-1004414 and by a Fulbright-Hays Research Fellowship for work at the Peace Research Institute, Oslo (`http://www.prio.no`). Addresses for authors: Leetaru : Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street, MC-493, Champaign, IL 61820-6211 USA; Schrodt: Department of Political Science, Pennsylvania State University, University Park, PA 16802 USA. A link to the current version of the GDELT data set can be found at `http://eventdata.psu.edu/data.dir/GDELT.html`

Abstract

GDELT—Global Data on Events, Location and Tone—is a new CAMEO-coded data set containing more than 200-million geolocated events with global coverage for 1979 to the present. The data are based on news reports from a variety of international news sources coded using the TABARI system for events and additional software for location and tone. The data is freely available and we expect to provide daily updates. This paper describes the news sources and some of their characteristics, the various processing steps that are used in generating the data, some comparisons with the KEDS Levants/Reuters and ICEWS/Asia data sets, and some visualizations. We conclude with an outline of planned enhancements to the data in the near future: these include recoding with new WordNet-enhanced dictionaries, the extension of the CAMEO coding to incorporate codes for financial events, disease outbreaks and natural disasters, and the development of an open-source Python-based successor to TABARI which will use parsed input from existing natural language processing tools.

Due to extensive graphics, the .pdf file for this paper is 16Mb, which exceeds the 4Mb limit of the ISA paper server. A copy of the paper can be downloaded from a link at `http://eventdata.psu.edu/papers.dir/automated.html`

1 Introduction

Political event data have had a long presence in the quantitative study of international politics, dating back to the early efforts of Edward Azar's COPDAB [Azar, 1980] and Charles McClelland's WEIS [McClelland, 1976] as well as a variety of more specialized efforts such as Leng's BCOW [Leng, 1987], though these efforts came to a standstill in the 1980s when funding for the large and costly human coding efforts—much of this provided by the U.S. Department of Defense Advanced Research Projects Agency—came to an end. Nonetheless, by the late 1980s, the NSF-funded[1] *Data Development in International Relations* project [Merritt et al., 1993] had identified event data as the second most common form of data—behind the various Correlates of War data sets—used in quantitative studies [McGowan et al., 1988]. The 1990s saw the development of two practical automated event data coding systems, the NSF-funded KEDS [Gerner et al., 1994, Schrodt and Gerner, 1994] and the proprietary VRA-Reader (`http://vranet.com`; King and Lowe 2004) and in the 2000s, the development of two new political event coding taxonomies—CAMEO [Gerner et al., 2009] and IDEA [Bond et al., 2003]—designed for implementation in automated coding systems. By the 2000s, with the decline of inter-state war, most event data studies shifted to the study of internal conflict, with the major project during this period being the $37-million U.S. Defense Advanced Research Projects Agency (DARPA) Integrated Conflict Early Warning System (ICEWS; O'Brien 2010).

These efforts built a substantial foundation for event data. By the mid-2000s, virtually all refereed articles in political science journals used machine-coded, rather than human-coded, event data. However, the overall investment in machine coding technology remained relatively small. This situation changed dramatically with the DARPA-funded Integrated Conflict Early Warning System project, which invested substantial resources in event data development using automated methods. The key difference between the ICEWS event data coding efforts and those of earlier NSF-funded efforts was the scale. As O'Brien—the ICEWS project director—notes,

> ...the ICEWS performers used input data from a variety of sources. Notably, they collected 6.5 million news stories about countries in the Pacific Command (PACOM) AOR [area of responsibility] for the period 1998-2006. This resulted in a dataset about two orders of magnitude greater than any other with which

[1] In view of recent developments in the U.S. Senate, we are placing more emphasis in this text on the degree to which this work has been possible due to funding from the NSF—predominantly the beleaguered NSF Political Science program—than might normally be the case. Thank you for your understanding; perhaps it will be contagious.

we are aware. These stories comprise 253 million lines of text and came from over 75 international sources (AP, UPI, and BBC Monitor) as well as regional sources (*India Today, Jakarta Post, Pakistan Newswire,* and *Saigon Times*).

While the original objective of ICEWS was conflict forecasting in Asia, the dependence of its most successful forecasting models on event data caused the program to morph into the production of a global event data set for 1996 to 2012, coded with the CAMEO event scheme and a customized sub-state actor scheme. Unfortunately, while the ICEWS project originally suggested that this data would be released for general use, it now appears to have disappeared into the classified world and there are no indications at present that the data will be available for use outside the U.S. government. Suggesting that the ICEWS models and data are proving to be very useful.

Fortunately, the availability of news texts on the web, along with various NSF-funded open source efforts at coding software and dictionary development, means that unlike the situation in the 1980s, it is possible to produce global data without the necessity of large-scale financial support. This paper will describe GDELT—Global Data on Events, Location and Tone—a new CAMEO-coded data set containing more than 200-million geolocated events with global coverage for 1979 to the present. The data are based on news reports from a variety of international news sources coded using the open-source TABARI system for events and additional software for location and tone. The data will be freely available to researchers at an NSF-funded server we are in the process of setting up at the University of Texas/Dallas—this may be operational by the time this paper is presented, and if not, shortly thereafter—and when fully operational, we expect this system will provide daily updates as well as user-friendly subletting capabilities. This paper describes the news text sources and some of their characteristics, the various processing steps that are used in generating the data, some comparisons with the KEDS Levants/Reuters and ICEWS/Asia data sets, some visualizations and concludes with an outline of future developments that we anticipate completing in the next six months.

2 Text Sources

Sources that were examined to identify events include all international news coverage from AfricaNews, Agence France Presse, Associated Press Online, Associated Press Worldstream, BBC Monitoring, Christian Science Monitor, Facts on File, Foreign Broadcast Information Service, United Press International, and the Washington Post. Additional sources exam-

ined include all national and international news coverage from the *New York Times*, all international and major US national stories from the Associated Press, and all national and international news from Google News with the exception of sports, entertainment, and strictly economic news.

The approximate distribution of the events over time is shown in Figure 1, which shows the total size of the files by year. Unsurprisingly, given the very substantial changes over the past two decade in both the international news environment and the availability of news on the web, this is anything but constant, and shows a dramatic increase since the beginning of the twenty-first century.

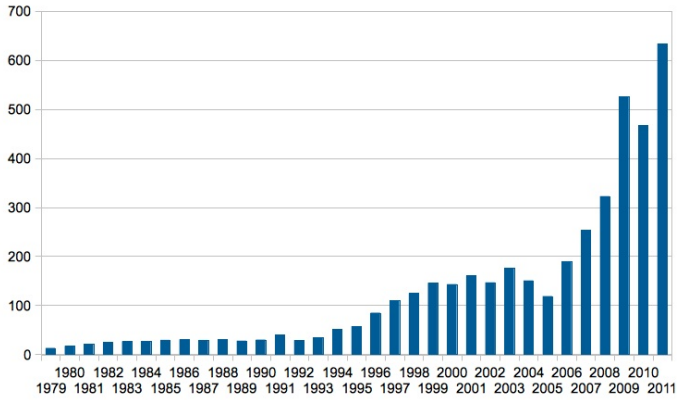


Figure 1: Distribution of GDELT events over time, Mb per year

In this section, we will look at three of the most important sources—AFP, AP and Xinua—in detail, both in terms of the availability of the data, the focus by country, and the variations in the level of coverage over time.

2.0.1 Agence France Presse

Agence France Presse is one of the largest news agencies in the world and is the largest in France. It is also one of the primary sources used by Western intelligence services to monitor the continent of Africa (Leetaru, 2010). LexisNexis describes the newswire in its Source Information directory as follows:

> Agence France Presse is the world's oldest news agency. Based in France, with staffers and stringers in 129 countries, AFP offers a unique perspective on the world's news. AFP's Europe coverage is outstanding, its reporting from Africa

> is renowned and its Latin American correspondence comprehensive. AFP also covers the Middle East, Asia and the Pacific Rim.

LexisNexis coverage of the newswire does not begin until May 1991. An extensive manual review of the source suggested it did not include an overrepresentation of coverage of domestic French affairs, focusing instead primarily on international coverage; earlier work by Phillip Huxtable came to a similar conclusion in comparing AFP and the English-language Reuters coverage of Ango- and Franco-phone West Africa. This suggested there was no need to incorporate additional filtering to specifically remove articles discussing French affairs. In addition, Agence France Presse articles occasionally quote French governmental officials on their views towards an emerging situation, which would result in many relevant articles being discarded if keyword-based filtering was used to remove all articles mentioning France or French. While the newswire contains SECTION() metadata tags used to identify the major news desks such as sports and financial news, these are not always properly applied. In addition, major sporting or financial news is often treated as general news, rather than being tagged under the appropriate section heading. An extensive manual review of a random selection of articles from each month was used to develop a lexicon of sports and financial-related keywords most commonly used in articles not properly tagged with the appropriate SECTION() tag. Thus, the following Boolean query was used to retrieve all articles from the Agence France Presse English file in LexisNexis:

> NOT section(sports) AND NOT section(financial) AND NOT golf AND NOT baseball AND NOT football AND NOT basketball AND NOT tennis AND NOT cycling AND NOT cricket AND NOT rugbyu AND NOT volleyball AND NOT "formula one" AND NOT subject(sports) AND NOT subject(financial results) AND NOT subject(economic news) AND NOT subject (stock indexes) AND NOT industry(stock indexes)

Figure 2 plots the total number of articles per month available in the LexisNexis archive of the newswire, showing that the newswire underwent steady growth through a peak in March 2001 and steadily decreased its output over the subsequent decade through mid-2010. It has largely remained constant at an average of around 8,000 articles a month over the last three years. There are also several outages visible in the first two years of its appearance in LexisNexis, which is noted on its LexisNexis Source Information page. In all, LexisNexis records 2,135,896 articles totaling 661,009,337 words through September 2012, averaging around 309 words per article.

Table 1 shows the top 25 countries most frequently discussed by Agence France Presse, ordered by the percent of all articles published by the service through September 2012 that mentioned each country. Here, any mention of the country or any city or other geographic landmark within the country was counted. Overall, there is a clear emphasis by Agence France Presse on Europe and the Middle East, with a particular focus on the United States

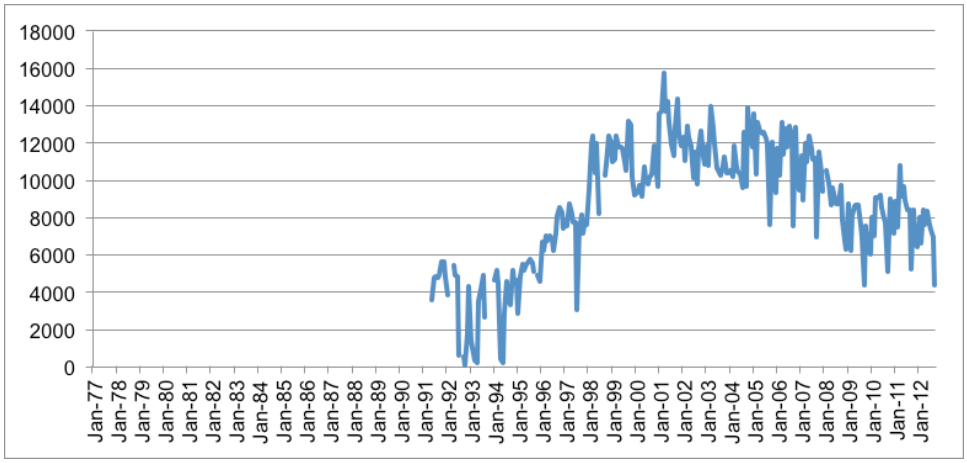


Figure 2: Articles per month Agence France Presse

and Great Britain.

2.0.2 Associated Press

The Associated Press is one of the largest news agencies in the world, operating 243 bureaus across the world. Unlike Agence France Presse, the Associated Press is operated as a cooperative, in which any story published by a member news agency is automatically redistributed and available for any other member to publish. LexisNexis describes the newswire in its Source Information directory as follows:

> Founded in 1848, and now delivering news and photos in over 100 countries, The Associated Press sees itself as the oldest and largest news service in the world. The AP is a nonprofit cooperative (i.e., a member-owned organization) with its roots in the newspaper industry. Regular members of the AP are obligated to report exclusively to the AP news that breaks locally, but might be of interest to the media elsewhere in the U.S. or overseas. This system gives the AP a news gathering reach well beyond what would be possible with only its staff resources. Coverage includes international news, national news (other than Washington-datelined stories), Washington news (only stories of national interest), business news, and sports.

The Associated Press newswire contains a wide assortment of news that heavily emphases domestic United States events, including local and regional newspaper coverage. Beginning in December 1978 the newswire added SECTION() metadata tags that allow the filtering of

Table 1: Top 25 countries by percent of all Agence France Presse articles mentioning that country

Country	All Articles
United States	10.44
United Kingdom	4.79
France	4.31
Russia	3.56
China	3.41
Israel	3.10
Iraq	3.08
Germany	2.72
Japan	2.36
India	1.78
Iran	1.73
Afghanistan	1.73
Pakistan	1.64
Italy	1.57
Egypt	1.33
Australia	1.31
Indonesia	1.25
Spain	1.22
Turkey	1.21
South Korea	1.15
Belgium	1.15
Syria	1.07
Canada	1.03
Saudi Arabia	0.98
Lebanon	0.98

coverage to just national or international stories (prior to this date there was no choice but to download all coverage). The newswire also has a special designation of top news used to identify major breaking or important stories regardless of their geographic focus. Thus, the following Boolean query was used to retrieve Associated Press coverage from LexisNexis:

"top news" or section(international)

Figure 3 plots the total number of articles per month available in the LexisNexis archive of the newswire, reflecting the far lower volume of coverage compared with Agence France Presse. The sharp drop in coverage volume between November and December 1978 reflects the introduction of the new SECTION() metadata tag that allowed for retrieving just international articles. While the newswire underwent steady growth during the late 1990s, it has experienced a decade-long decline in its international coverage, stabilizing at around 1,600 articles per month over the last three years. In all, LexisNexis records 944,483 articles totaling 358,398,400 words through September 2012, averaging around 379 words per article. Table 2 shows the top 25 countries in terms of the relative percentage of all Associated Press coverage during this time that mentioned each. As with Agence France Presse, there is a strong emphasis towards European and Middle Eastern countries and a similar emphasis on French coverage.

Those familiar with the Associated Press will likely question why the primary Associated Press newswire was used here, rather than the specialty Associated Press Worldstream newswire, which is exclusively focused on international news coverage. LexisNexis does in fact offer an archive of the Worldstream service that begins in October 1993 that is very comparable in terms of daily coverage volume to Agence France Presse. However, for unknown reasons the LexisNexis archive of Worldstream ends abruptly in July 2010, with coverage past this date exclusively carrying sporting results.

2.0.3 Xinhua

Xinhua is the official news agency of the People's Republic of China and the largest news service in the country, operating 107 bureaus around the world. While it still retains its official role in promulgating the views and statements of the Communist Party, it has vastly expanded since its founding in the 1931 towards a general-purpose global news service competing with services like Reuters (Troianovski, 2010). LexisNexis describes the newswire in its Source Information directory as follows:

Xinhua is the authoritative source for information on Chinese government affairs,

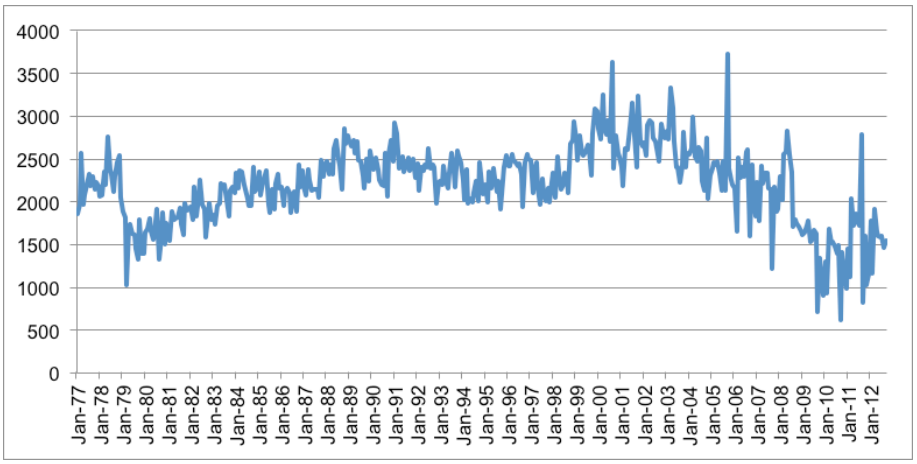


Figure 3: Articles per month Associated Press

Table 2: Top 25 countries by percent of all Associated Press articles mentioning that country

Country	All Articles
United States	13.35
United Kingdom	5.32
Russia	4.51
France	3.39
Israel	3.36
Germany	2.74
Iraq	2.71
China	2.27
Japan	1.94
Iran	1.87
Italy	1.87
Egypt	1.50
Afghanistan	1.44
Lebanon	1.44
Canada	1.27
Pakistan	1.26
India	1.26
Spain	1.15
Syria	1.14
West Bank	1.12
Saudi Arabia	1.05
Mexico	1.04
South Africa	0.98
Poland	0.97
Turkey	0.97

> economic performance and Chinese views on world affairs. All Western news correspondents in Beijing rely on Xinhua's English-language news report to keep abreast of Chinese affairs. The agency reports on Chinese affairs, including the economy, industry, trade, agriculture, sports and culture. Coverage includes diplomatic changes and extensive international reporting often from Africa or the Middle East. Xinhua also provides useful coverage of non-Chinese Asia.

As its description above suggests, Xinhua is extensively focused on domestic Chinese news, which would heavily overemphasize China over other countries. Through manual review of a random selection of articles from each month, it was determined that adding in exclusion keywords to drop those articles mentioning either China or Chinese removed domestic coverage with a minimal false positive rate. Unlike Agence France Presse, Xinhua coverage of international events uses quotes from Chinese officials far more sparingly, meaning this filtering criterion has a minimal impact on international coverage. Xinhua also has a dedicated financial newswire called Xinhua Economic News Service, separating Xinhua's extensive coverage of the financial markets from the Xinhua General News Service newswire used here. It does not, however, offer the SECTION() tags used with Agence France Presse and Associated Press coverage to filter out sports-related coverage, necessitating the use of additional keyword filters. Thus, the following Boolean query was used to retrieve Xinhua coverage from LexisNexis:

> NOT china AND NOT Chinese AND NOT olympic AND NOT snooker AND NOT boxing AND NOT hockey AND NOT marathon AND NOT motorcycling AND NOT soccer AND NOT handball AND NOT cycling AND NOT tennis AND NOT world cup AND NOT basketball AND NOT wrestling match AND NOT wrestling score AND NOT iceskating

Figure 4 plots the total number of articles per month available in the LexisNexis archive of the newswire. The near-tripling of coverage between December 1998 and November 1999 reflects the US involvement in Iraq during this period, which attracted a singularly massive volume of coverage from Xinhua. The service also has two major outage periods in LexisNexis, from April 1995 to June 1996 (inclusive) and April 2008 to October 2008 (inclusive), so those are removed from consideration for all analyses. In all, LexisNexis records 1,699,442 articles totaling 332,043,292 words through September 2012, averaging around 195 words per article. Table 3 shows the top 25 countries in terms of the relative percentage of all Associated Press coverage during this time that mentioned each. As with Agence France Presse and the Associated Press, there is a strong emphasis towards European and Middle Eastern countries.

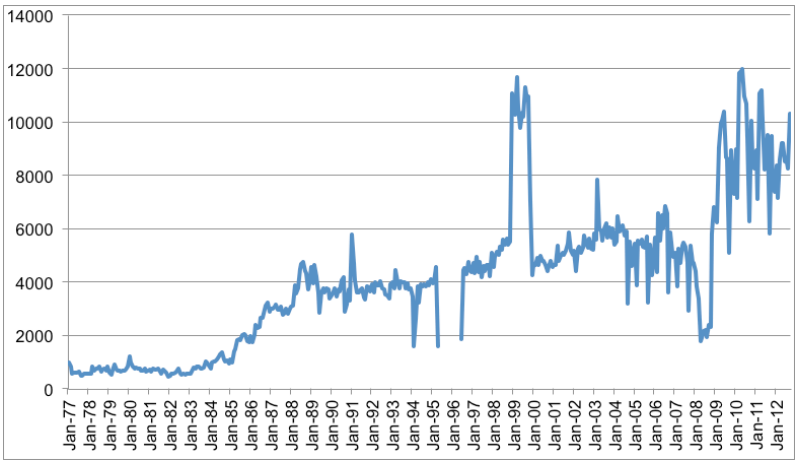


Figure 4: Articles per month Xinhua

Table 3: Top 25 countries by percent of all Xinhua articles mentioning that country

Country	All Articles
United States	10.76
Israel	3.81
United Kingdom	3.34
Russia	3.27
Iraq	2.96
France	2.43
Japan	2.32
Pakistan	2.01
India	1.98
Egypt	1.94
Iran	1.94
Germany	1.92
Afghanistan	1.88
Thailand	1.53
Philippines	1.44
South Africa	1.42
Australia	1.41
Turkey	1.37
Indonesia	1.34
Syria	1.26
Lebanon	1.14
Nigeria	1.08
West ank	1.07
Kenya	1.07
Italy	1.06

2.0.4 Comparing the Sources

This section will briefly compare some of the characteristics of the three news sources in order to tease apart their mutual differences and explore whether predictive features found in one source are universal across the others; additional details can be found in Leetaru [2013]. In total across the three sources, 4,779,821 articles were processed in the course of this dissertation, totaling 1,351,451,029 words. Figure 5 shows the Z-scored (standard deviations from mean) plots for all three sources overlapping their relative growth and decay patterns. Figure 6 shows the combined monthly article volume across the three sources demonstrating in particular the significant mutual growth during the 1990s. The three sources, however, are poorly correlated in their temporal profiles. Even restricting the analysis to only overlapping periods of time, the monthly coverage volume of Agence France Presse has a Pearson correlation of $r = 0.27$ with Xinhua and $r = 0.35$ with Associated Press, while the Xinhua and the Associated Press are correlated at $r = 0.03$. While weak, Agence France Presse is correlated with the other two sources at $p < 0.0005$, indicating high statistical significance, with Agence France Presse and the Associated Press being the only two sources not to have a statistically meaningful correlation. In terms of geographic emphasis, the three sources are tightly aligned, with Agence France Presse being correlated at $r = 0.97$ in terms of the relative percentage of eachs coverage dedicated to each country, while Agence France Presse is correlated at $r = 0.94$ with Xinhua. Xinhua and the Associated Press are correlated at $r = 0.95$. All three are therefore at the highest level of statistical significance ($p < 0.0005$), indicating there are no significant differences in their respective geographic focus.

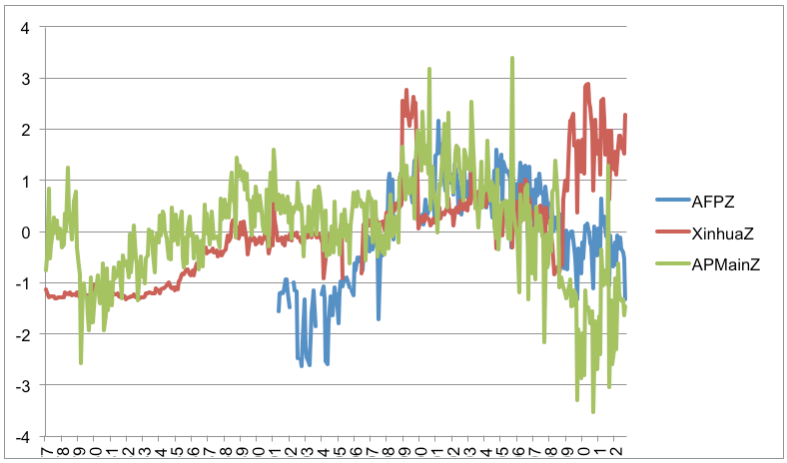


Figure 5: Z-scored articles per month comparing Agence France Presse, Associated Press, and Xinhua

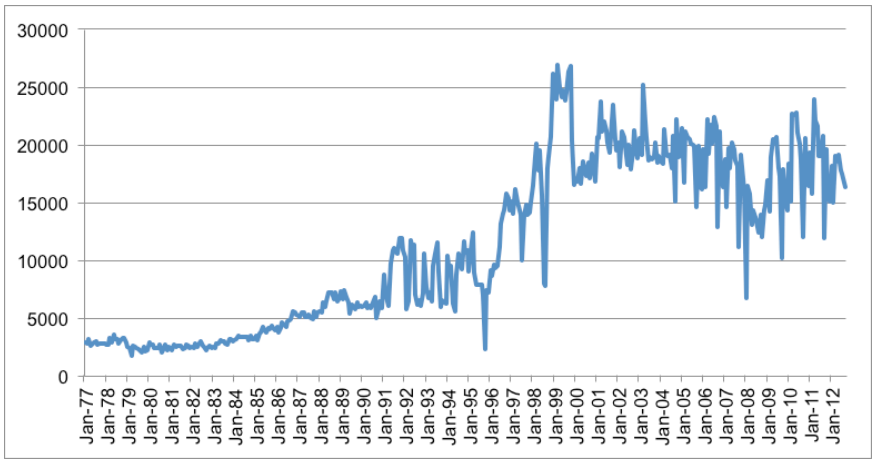


Figure 6: Combined articles per month across Agence France Presse, Associated Press, and Xinhua

2.0.5 Additional Post Filtering

Despite careful construction of the queries used to retrieve each newswire from LexisNexis, including extensive manual review of random selections of content to develop exclusion keywords, a non-insignificant volume of sports and financial coverage was still retained. Such coverage can create complications for the event coding process in that it often contains language very similar to that used to describe violent political events, such as two sports teams battling it out or a companys stock price "under siege" [Schrodt and Van Brackle, 2013]. Thus, after all matching articles were downloaded from each newswire, a second extensive manual review was performed across the combined content pool to filter out any remaining sports or economic-related coverage. While this filter may eliminate certain economic-related articles which might reflect or drive public opinion (such as an economic boom or bust), incorporating this filter dramatically reduced the false positive rate of TABARI. Thus, the following Boolean query was applied in a post-processing stage after the content was downloaded, but before it was made available for secondary processing.

NOT boxing AND NOT hockey AND NOT marathon AND NOT motorcycling AND NOT soccer AND NOT handball AND NOT cycling AND NOT tennis AND NOT worldcup AND NOT world cup AND NOT basketball AND NOT wrestling match AND NOT wrestling score AND NOT icestaking AND NOT ice staking AND NOT skiing AND NOT football AND NOT coach AND NOT hockey AND NOT box office AND NOT snooker AND NOT cricket AND NOT game console AND NOT gaming console AND NOT tv show AND NOT bond market AND NOT currency trade AND NOT closed up AND NOT closed down AND NOT industrial average AND NOT nasdaq AND NOT dow jones AND NOT halftime AND NOT half time AND NOT the game AND NOT stocks declined AND NOT market declined AND NOT inflation AND NOT interest rate AND NOT regional growth AND NOT car sale AND NOT truck sale AND NOT midsize car AND NOT inflation AND NOT singer AND NOT teammate AND NOT team mate AND NOT freethrow AND NOT free throw AND NOT show times AND NOT athletic AND NOT free throw AND NOT touchdown AND NOT the season AND NOT rebounds AND NOT quarterback AND NOT point guard AND NOT fourth quarter AND NOT on the road

AND NOT season high AND NOT diet AND NOT title bid AND NOT mixed doubles AND NOT bowl game AND NOT retail price AND NOT book review AND NOT garden AND NOT goalkeep AND NOT goal keep AND NOT mega million AND NOT megamillion AND NOT mega-million AND NOT lottery game AND NOT lottery winner AND NOT ticket sale AND NOT lottery jackpot AND NOT baseball AND NOT golf AND NOT growth outlook AND NOT the dollar AND NOT bank index AND NOT nfl AND NOT nhl AND NOT nba AND NOT sports AND NOT championship AND NOT entertainment

This leads to the following processing pipeline used to construct the event database used here:

1. All relevant content from each newswire is downloaded from LexisNexis Academic Universe using the source-specific query.
2. Each article is subjected to the additional post-processing Boolean query to drop remaining sports- and financial-related news coverage.
3. Each article is subjected to fulltext geocoding from Leetaru [2012] to identify and disambiguate all geographic references contained in each article.
4. The TABARI system is applied to each article in full-story mode to extract all events contained anywhere in the article and the TABARI geocoding post-processing system is enabled to georeferenced each event back to the specific city or geographic landmark it is associated with.
5. The final list of events for each newswire is internally deduplicated. Multiple references to the same event across one or more articles from the same newswire are collapsed into a single event record. To allow the study of each newswire individually, events are not deduplicated across newswires (externally deduplicated).

In all, there were 28,877,172 events identified and coded by TABARI from the three news wires: 14,433,748 from Agence France Presse, 7,811,104 from the Associated Press, and 6,632,320 from Xinhua.[2] On average, there were 7 events per article in Agence France Press, 8 from Associated Press, and 4 from Xinhua. However, since the three sources have very different average article lengths, when calculating the average words per event, the results are far more even: 46 words per event on average for both Agence France Presse and the Associated Press, and 50 words per event for Xinhua.

Figures 7, 8 and 9 show the total number of events per month for each news source in the four Quad Classes of Verbal Cooperation, Material Cooperation, Verbal Conflict, and Material Conflict. The gap in Associated Press events for 1992 is due to a technical error with the content downloaded for that year that prevented it from being properly processed, this has

[2]The complete GDELT data set incorporates additional stories from other sources.

been corrected in the final data set. Both Agence France Presse and Xinhua have very similar volumes of Material Cooperation and Verbal and Material Conflict events, making it hard to distinguish the relative change over time in those three categories, while their Verbal Cooperation events have clear temporal signatures. Associated Press coverage, seen in Figure 8, however, has strong stratification among the four classes, making it clear that they are closely aligned. Indeed, the four series are correlated at between $r = 0.80$ and $r = 0.97 (p < 0.0005)$ for all three news wires. Finally, Table 4 shows the relative breakdown of all events into the four Quad Classes for each newswire. It is clear that Verbal Cooperation events are the most common, followed by Material Conflict, Verbal Conflict, and Material Cooperation. In all, across the three sources, 60.56% of events were Verbal Conflict, 17.34% were Material Conflict, 13.12% were Verbal Conflict, and 8.99% were Material Cooperation.

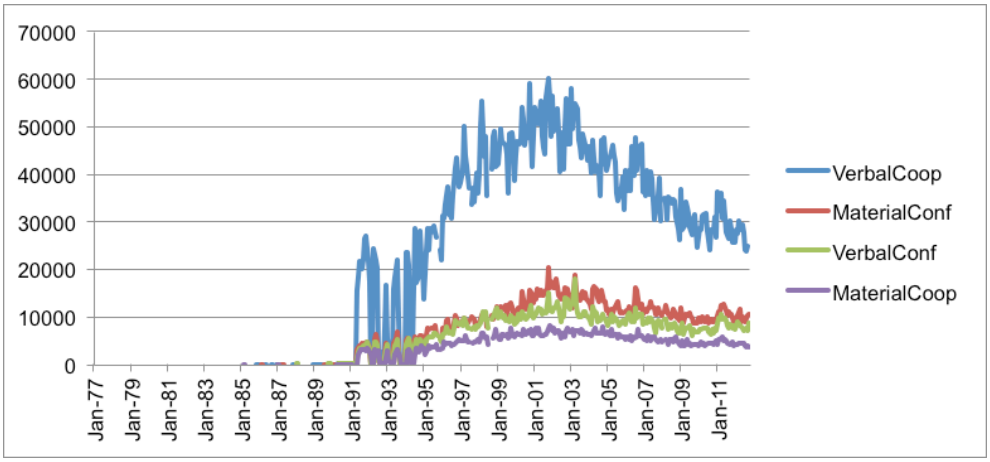


Figure 7: Agence France Presse events per month by Quad Class

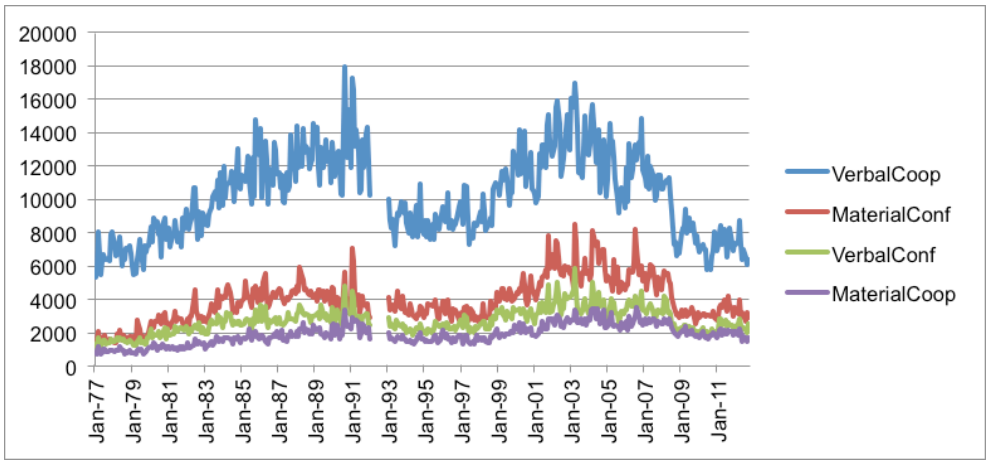


Figure 8: Associated Press events per month by Quad Class

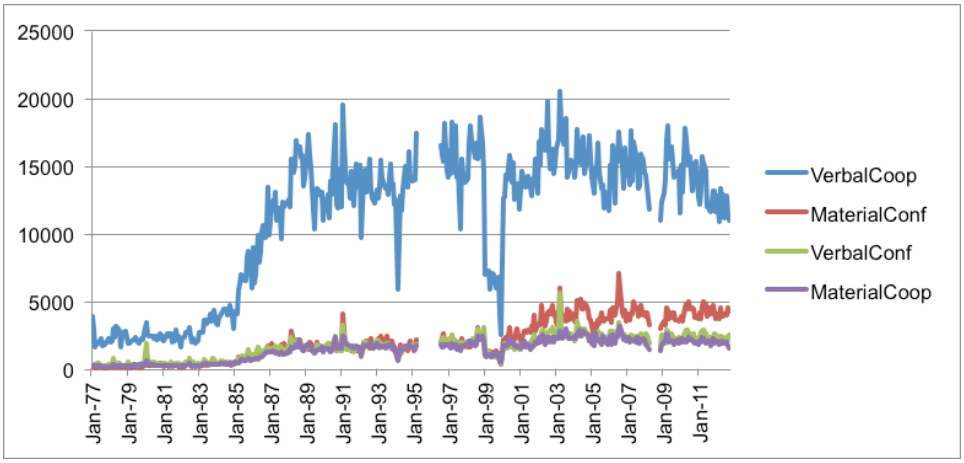


Figure 9: Xinhua events per month by Quad Class

Table 4: Breakdown by percent of all events in each newswire in each Quad Class

	AFP	AP	Xinhua
Verbal Cooperation	60.35	54.63	68.02
Material Conflict	17.37	20.75	13.22
Verbal Conflict	13.75	14.45	10.16
Material Cooperation	8.53	10.17	8.60

3 Automated Coding Engines

In the field of event data, multiple independent tests [King and Lowe, 2004, Schrodt and Gerner, 1994] have shown that machine coding is comparable in accuracy to human coding. Furthermore, the human coding accuracy in some of those tests is quite low: King and Lowe [2004] find the accuracy on the individual VRA event codes alone (Table 2, pg 631)—not the complete record with source and target identification—is in the range 25% to 50% for the detailed codes and 55% - 70% for the major categories. Similarly, Mikhaylov et al. [2012] show that the reliability of the human coding in the widely-used Comparative Manifestos Project is less than half what is commonly reported, and for some indicators drops as low as 25%; Ruggeri et al. [2011] show similar problems in the coding of governance events in UN peacekeeping.

An extensive recent body of psychological work—see Baumeister and Tierney [2011] for a popular treatment—indicates that the sustained decision-making required for human coding presents an almost perfect storm for inducing fatigue, inattention, and a tendency to use heuristic shortcuts. These psychological costs are far more deeply rooted than previously assumed, and can only be reduced, not eliminated, by improved coding protocols, training,

coder selection and supervision. The human brain was simply never intended for the tasks we impose on coders.

Improvements in computer software and hardware, meanwhile, have made the process of analyzing large bodies of text much more efficient, and the field has matured with the development of a common set of methodologies with well-tested characteristics. Automated coding is completely transparent, without the unreproducible subjective elements of human coding. Moreover, once the source texts have been prepared, recoding to account for new theoretical or technological components can be done quickly and efficiently.

Automated coding also opens the possibility of near-real-time (NRT) coding, which can scale to arbitrarily large sets of source texts. In the early phases of the ICEWS project, an implementation of TABARI on a small cluster computer reliably coded 26-million news reports in six minutes, resulting in about 3-million events. In contrast, sustained human coding projects, once one accounts for training, retraining, replacement, cross-coding, re-coding due to effects of coding drift and/or slacker-coders and so forth, usually code about six events per coder-hour and–like the labor requirements of a string quartet–have changed little over time. The arithmetic is obvious: 3-million events from six minutes of automated coding, or 500,000 labor-hours of manual coding, probably costing on the order of $10-million when labor and administrative costs are taken into effect.

3.1 Event Coding

3.1.1 TABARI

The core engine for the event coding was the open-source TABARI program. While this program has been in use for a number of years, the GDELT project provided an incentive for several enhancements

- We improved the ability of TABARI to automatically assemble codes from combinations of a named actor and an generic agent. For example "Philippine soldiers" will automatically generate the code `PHLMIL`, whereas "The Philippine Secretary of Agriculture" will automatically generate the code `PHLGOV`. Earlier dictionaries had done this directly, with separate dictionary entries for, say, "Australian police," "Cambodian police," "Chinese police" and so forth. The new system is both faster in terms of the dictionary size and much more efficient. This allows the coding of both generic agents such as "police", "soldiers", "demonstrators" and the like, as well as named individuals where we have the title in the dictionary but not the individual person. In support

of this new facility, we also increased the size of the *.agents* dictionary considerably based on *WordNet*.

- The TABARI `CONVERT AGENTS` facility generalizes this further and allows agents which do not have an associated actor to be converted to actors and their agent codes (rather than actor+agent codes) are used. This is allows a sentence such as

  Students and police fought in the Egyptian capital

  to be coded as

  `EDU fought COP`

  or the sentence

  Sudanese students and police fought in the Egyptian capital

  to be coded as

  `SUDEDU fought COP`

  This allows coding where the location of the event can be determined by processing with other programs. The use of this option increased the number of coded events by about 22%.

- A new facility was added to TABARI that made the source and target actors optional: this allows both for events such as general statements where there is not an explicit target, and also the coding of situations where an event is implied by the structure of the sentence, but the system cannot find the appropriate source or target in the dictionaries. Sentences coded in this manner contain a default code so that they can be easily eliminated at a later processing stage if a researcher does not want to include this information, but are useful in aggregations where the identity of the [typically] target is not important.

3.1.2 Dictionaries

The TABARI system has very extensive open-source dictionaries for the identification of political actors. Central to these is the 32,000-line `CountryInfo.txt` (`http://eventdata.psu.edu/software.dir/dictionaries.html`), a general purpose file intended to facilitate natural language processing of news reports and political texts. This covers about 240 countries and administrative units (e.g. American Samoa, Christmas Island, Hong Kong, Greenland); fields include adjectival forms and synonyms of the country

name, the capital city and cities with populations over 1-million, regions and geographical features (*WordNet*meronyms), leaders from `http://rulers.org` and members of government from the *CIA World Leaders* open database.

`CountryInfo.txt` was supplemented by a number of lists of actors such as MNCs, NGOs and IGOs, and militarized non-state actors such as al-Qaeda and the Lords Resistance Army. Under NSF funding, the CAMEO actor coding system has been enhanced with a *very* extensive religion typology, the CAMEO Religious Coding Scheme (CAMEORCS), which has a hierarchical coding of about 1,500 religious denominations, and an ethnic group coding scheme with about 650 ethnic groups. The formal names of these have been implemented as TABARI dictionaries, though we have not had the opportunity to evaluate how frequently this translates into actual actor codes.

As a consequence of these enhancement, the *.actors* dictionaries now have around 60,000 entries, compared to the 1,000 or so entries typical in earlier KEDS work; this is further supplemented with about 1,500 agents (common nouns), which can also be treated as actors either in combination or alone.

The verb coding dictionary was the standard one we have been using since about 2010—this was a point when regular noun and verb forms were added to TABARI—which proved remarkably robust in a number tests we did in the early phases of ICEWS. As noted in the conclusion, we will soon be shifting to a new dictionary which employs a extensive set of WordNet synonyms both in the verbs themselves and also in common phrases where synonym sets might be encountered, such as the names of monetary currencies and the names of weapons. We have identified about a dozen such categories and have largely incorporated these into the dictionaries, but this was not used in this version of GDELT.

3.2 Location and Tone

The raw TABARI output files are geolocated by a post-processing system. The first system takes the verb and actor word sentence offsets (that is, the identity of the word in the sentence that was the start of each actor and verb mention) and converts them back into their locations in the original article. This is then cross-walked against the list of all geographic locations found in the text, and the closest location to each of the actor and verb mentions is assigned to it. While this is quite primitive, it works surprisingly well and was the approach used to map Wikipedia.[3] The "tone" of each article is measured using the tonal algorithm from

[3] `http://www.dlib.org/dlib/september12/leetaru/09leetaru.html`

Shook et al. [2012].

3.3 "The Pipeline"

3.3.1 Pre-processing

- First the article is cleaned up, any textual URLs, phone numbers, email addresses, non-ASCII characters, and other material is removed.
- The article is then run through full-text geocoding that automatically identifies and disambiguates all geographic references in the text, resolving them to their centroid lat/long coordinates, using Leetaru (2012).[4]
- The article then goes through a preprocessing pipeline that creates four versions of the text:
  - *RAW*: The text is kept as-is with no change.
  - *CITYTOCOUNTRYNOPERSON*: All mentions of cities and other geographic landmarks are replaced with the name of the country the city/landmark is located in. Thus, "Soldiers marched in Cairo today" will be replaced with "Soldiers marched in Egypt today." While the TABARI *.actors* dictionary contains major city names, it contains only the largest cities, and a mention of soldiers attacking a small village will cause TABARI not to find a country affiliation for the victim actors, while this interpolation will replace that small village with its corresponding country name, causing TABARI to find a match. Person names remain as-is.
  - *PERSONTOCOUNTRY*: For each person's name found in the text, the closest location to the first mention of the person's name is then assigned to that person and all mentions of the person's name are then replaced with that location's country name. Thus, a mention of "Egyptian Minister of Foreign Affairs Mohamed Orabi attended the summit yesterday. While he was there, Orabi pledged support for..." the second sentence would be rewritten to "While he was there, Egypt pledged support for..." This is essentially a special geographically-centered form of entity dereferencing. This version dramatically improves matching of diplomatic events in which a political leader is referred to throughout an article, but the leader is not significant enough to have his or her own entry in the TABARI

[4] `http://www.dlib.org/dlib/september12/leetaru/09leetaru.html`

.actors list. In addition, all mentions of cities and other geographic landmarks are replaced with "CityName, CountryName."

– *FULL*: In this version, each person name is associated with its corresponding location as in *PERSONTOCOUNTRY*, but all mentions of the person's name are replaced with the form of "PersonName of CountryName" (adding the country name and inserting the word "of" between the person's name and the country name). All mentions of cities or landmarks are converted to the format "CityName, CountryName."

3.3.2 Coding

Each of the text versions above causes TABARI to catch different events and significantly improves TABARI's ability to capture complex events, since it essentially "rewrites" each sentence several different ways for the TABARI grammar to catch it: these rules make the sentence "TABARI-friendly."

- For each of the versions above, the text is processing through TABARI twice, once with the standard settings and once with `COMMA: OFF` set in the *.options* file. Through extensive testing, it appears that on more complex articles, this dual-pass coding process yields a substantial increase in the number of recovered events without yielding a measurable number of false positives.

- TABARI is designed to process only a single sentence at a time, so the above text is then converted to the per-sentence input format needed by TABARI . A special sentence-splitting algorithm is used that is robust to periods occurring in different contexts such as abbreviations and typographical errors, and multiple blank lines being used as paragraph separators, etc. In addition, an optimization process uses a fast surface scan of the sentence to determine whether it contains matches of at least one entry from the TABARI *.verbs* list. It does not attempt to determine if there would be an actual match based on a grammatical parse, but simply checks whether it would even be possible for the sentence to have a match. This is extremely fast and eliminates 70-80% or greater of all sentences, vastly speeding up the rest of the processing pipeline (both the input/output requirements of managing all of those sentences and TABARI's processing time).

- While TABARI ordinarily is used only to code the first sentence of each article, here it is used to code all sentences to ensure it codes all of the surrounding contextual

events. This vastly improves its recognition of complex situations as well in that while the lead sentence may be too grammatically complex in some cases for TABARI to process, those events will be repeated later in the text, so it ultimately codes all of the events. Events coded in the first lead paragraph of the article are marked using the RootEvent flag so that users can select only these events as desired. In addition, TIME SHIFTING is enabled to ensure that mentions of past events are coded with the proper date.

- Historically TABARI would code a blank actor if there was AGENT information available, but no actor (such as gunmen or police). The new CONVERT AGENT = TRUE functionality of TABARI is used here so that in the absence of an available actor, TABARI will code any available AGENT mentions as the actor instead. This yields an average of 22% additional events that were previously lost.

3.4 One-A-Day Filtering

Following the protocols used in most of the TABARI-based research, the major post-processing step is the application of a "one-a-day" filter, which eliminates any records that have exactly the same combination of date, source, target and event codes. This is designed to eliminate duplicate reports of events that were not caught by earlier duplicate news report filters. In our work on the Levant data set, this fairly consistently removes about 20% of the events; the effect on the ICEWS data may be somewhat higher due to the use of a greater number of sources.

In areas of intense conflict—where multiple attacks could occur within a single dyad in a single day—this could eliminate some actual events. However, these instances are rare, and periods of intense conflict are usually obvious from the occurrence of frequent attacks across a month (our typical period of aggregation), and do not require precise measures within a single day. Periods of intense conflict are also likely to be apparent through a variety of measures—for example comments, meetings with allies, offers of aid or mediation—and not exclusively through the attacks themselves.

Each record is then converted into a unique identifier key that concatenates the actor, action, and location fields. This is then checked against the list of all existing events in the database: if the event already exists, the previous event record has its $NumMentions$ and $NumSources$ fields updated accordingly, otherwise a new record is inserted into the database. Traditional TABARI event deduplication is performed at the country-day level, however in high-conflict

regions like Syria, it is necessary to incorporate the city-level information into the event record to preserve the locations of each riot and to separate riots in different parts of the country into distinct event records.

4 Visualizations

4.1 Quad Count Time Series for Selected Dyads

The figures in Table 5 shows quad-count time series plots[5] for four active dyads: Israel-Palestine, Israel-Lebanon, China-Taiwan and India-Pakistan. With the exception of the Israel-Lebanon case, the most apparent feature of these is the increase in the number of events over time; this scaling issue also masks significant variation in each series in the first half of the data.

4.2 District-level Conflict in Afghanistan

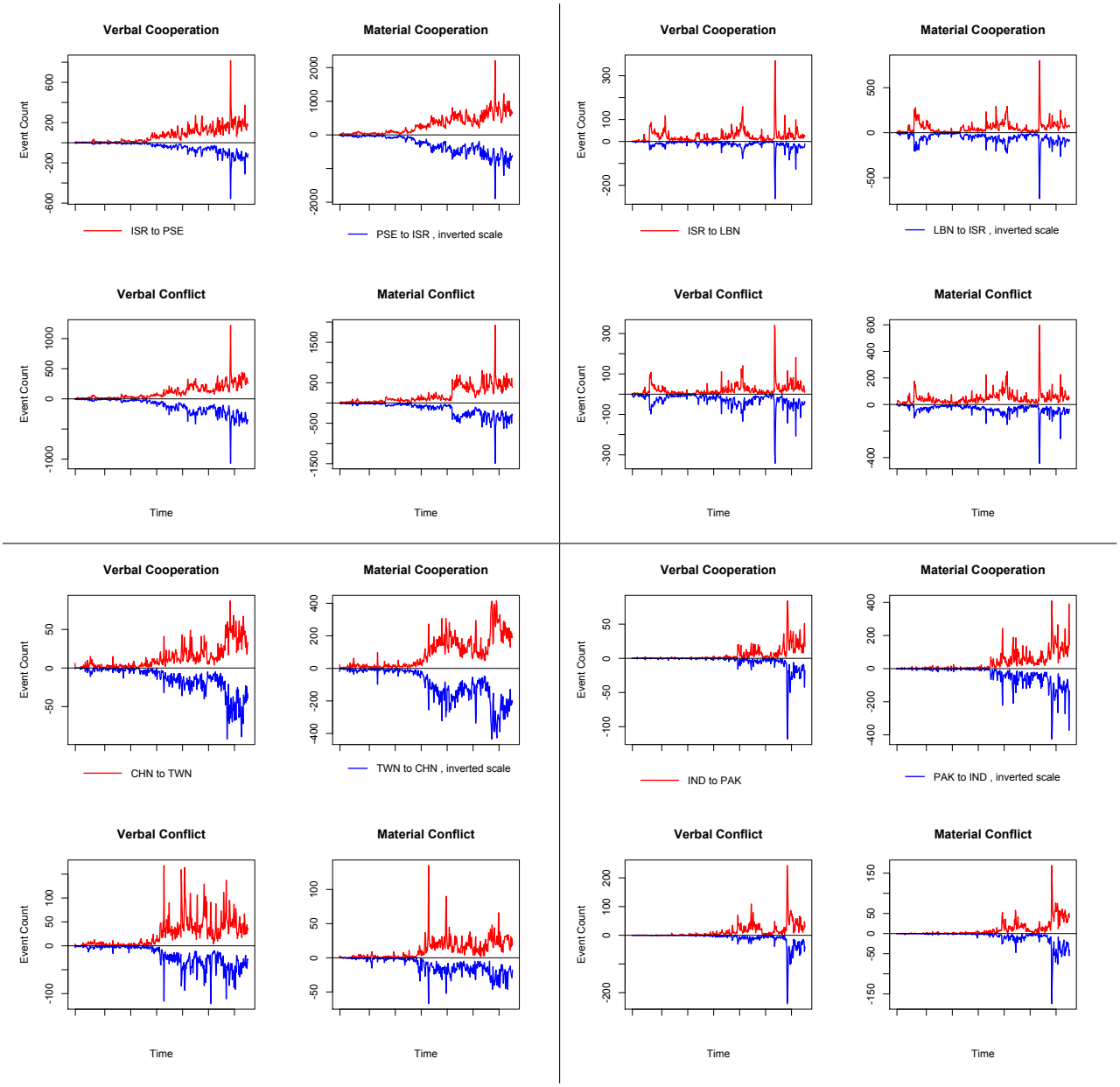
Yonamine [2013, chapt. 4] uses GDELT to develop models which predict district-level conflict in Afghanistan, finding that the more finely-aggregated district data produces more accurate predictions than province-level and national-level data. Figure 10 shows the distribution of the data over time: note that this looks similar to maps that have been produced using the Wikileaks data—the extent to which this can be used in research still very much in legal limbo—but was produced entirely from open-source, contemporaneous sources.

4.3 Comparision of GDELT and NGO-based data for City-level violence in Syria

Yonamine [2013, chapt. 1] compares city-level reports of violent events in GDELT and the level of violence reported by a Syrian-based NGO: these two sources are shown in Figures 11 and 12. Despite that fact that Syria is a particularly difficult conflict to cover—on a typical day, the number of deaths reported in the major international media are less than half those reported in an NGO source such as `syriashuhada.com`—the shape of the curves for Aleppo and Homs are almost identical between the two distinct sources.

[5]Python and R code for generating these is included in the "GDELT.reduced" package that can be downloaded at `http://eventdata.psu.edu/data.dir/GDELT.html`.

Table 5: Quad Count Time Series for Selected Dyads



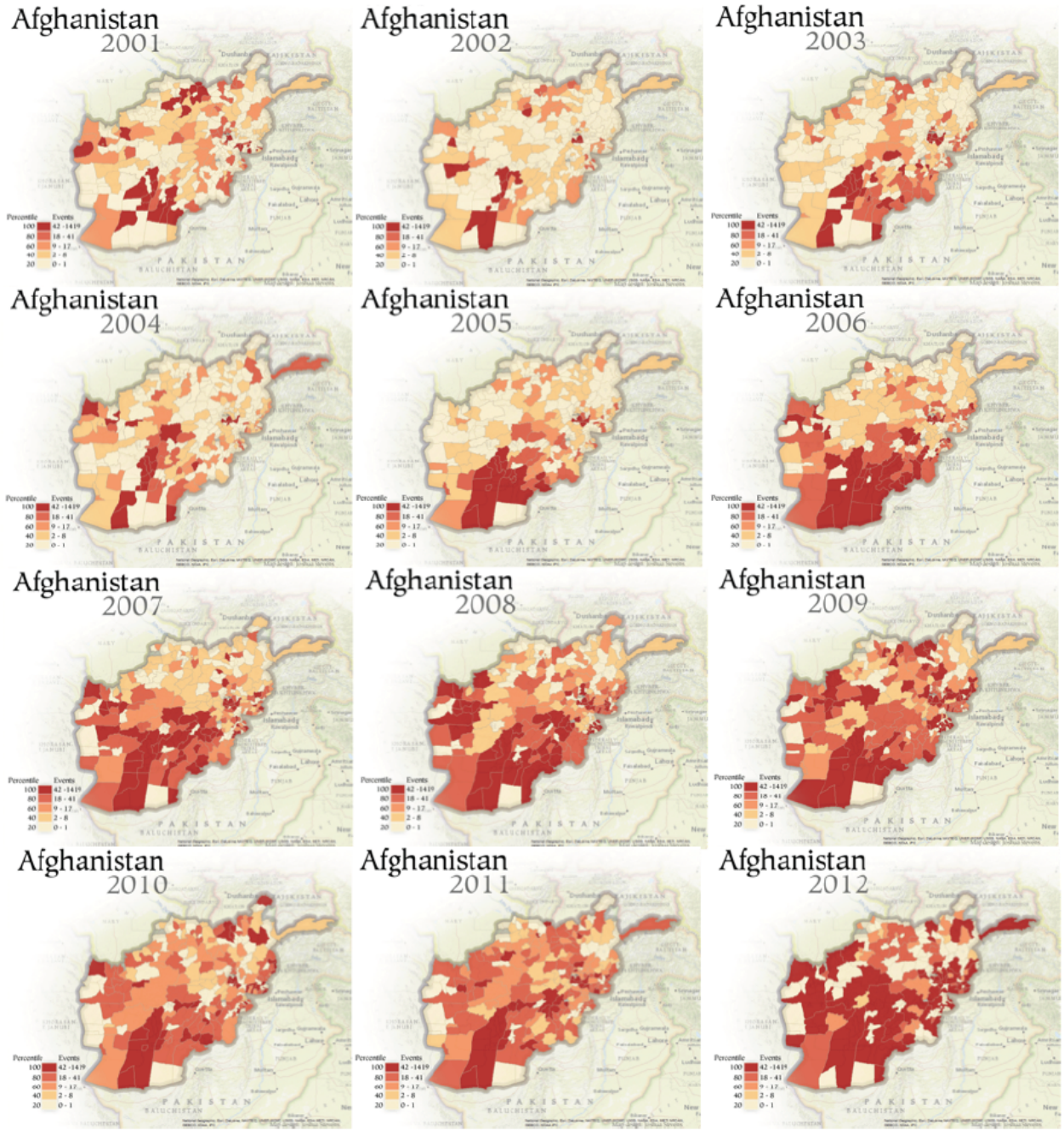


Figure 10: The Number of Material Conflict events per Afghani District from 2001 to 2012 (Yonamine 2013)

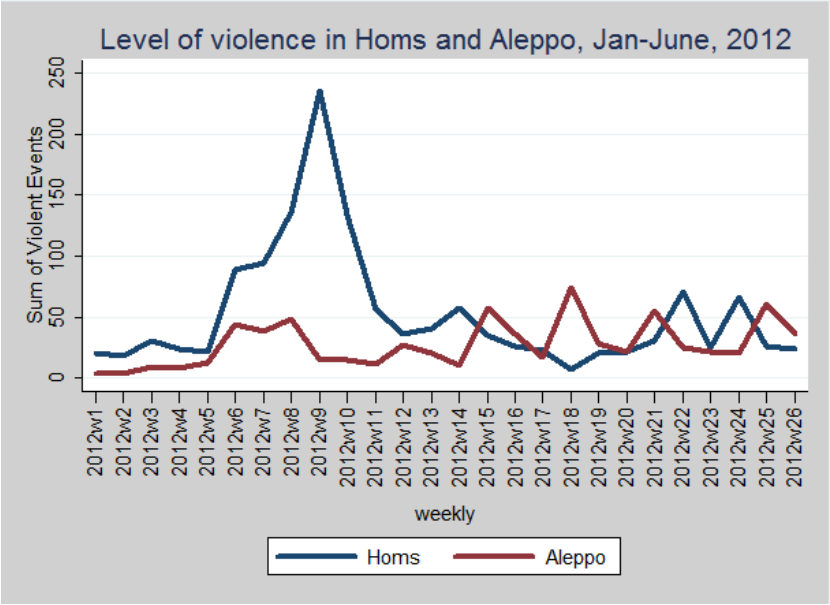


Figure 11: The Number of GDELT-derived Violent Events in Homs and Aleppo from January 2012 through June 2012 (Yonamine 2013)

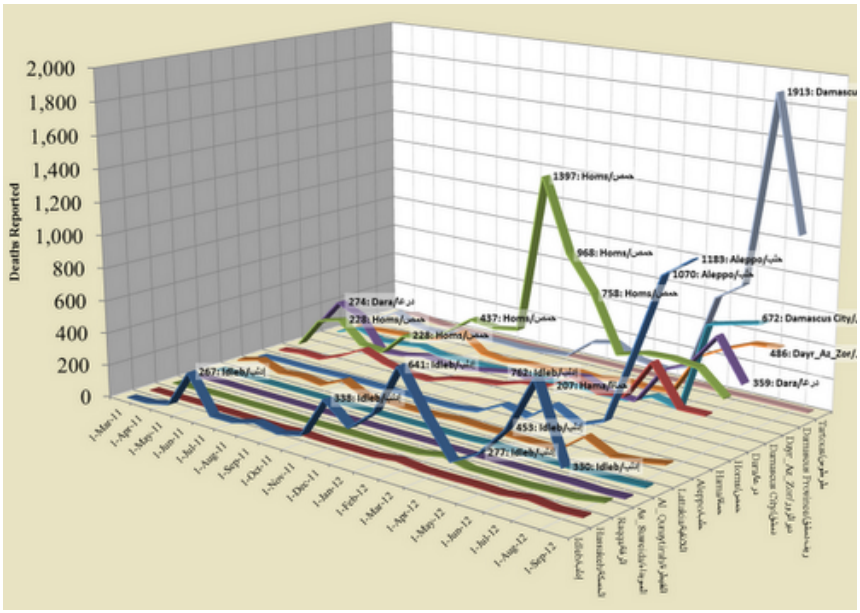


Figure 12: The Number of NGO-reported Violent Events in Homs and Aleppo from January 2012 through June 2012 (Yonamine 2013)

4.4 GDELT: 29 January 2011

Figures 13 through 15 show three different visualizations of the GDELT "world on 29 January 2011, in the middle of the Egyptian Spring.

Figure 13 uses modification of a special heatmapping process developed by the cybergis group (`http://www.psc.edu/media/training/XHPC2012/Culturomics.pdf`; Shook et al. [2012]). For each *Action* location it evaluates the total number of events at the point and the average Goldstein score [Goldstein, 1992] of all of those events. As such, it gives a rough heatmap showing the major areas of positive and negative events, emphasizing the underlying clusters. This technique shows the major groupings of positive and negative actions, filtering them from the background landscape of events.

Figure 14 shows the same day, but uses the raw event data. It sizes the dot by the $NumberMentions$ field; the colors are based on the Goldstein scale. For clarity, $NumberMentions$ is restricted to a maximum, so all major events are around the same size.

Figure 15 puts dots for the locations of both $Actor1$ and $Actor2$ and then draws a line between them colored by the average Goldstein score of the event. This shows which areas of the world are being connected to what other areas of the world.

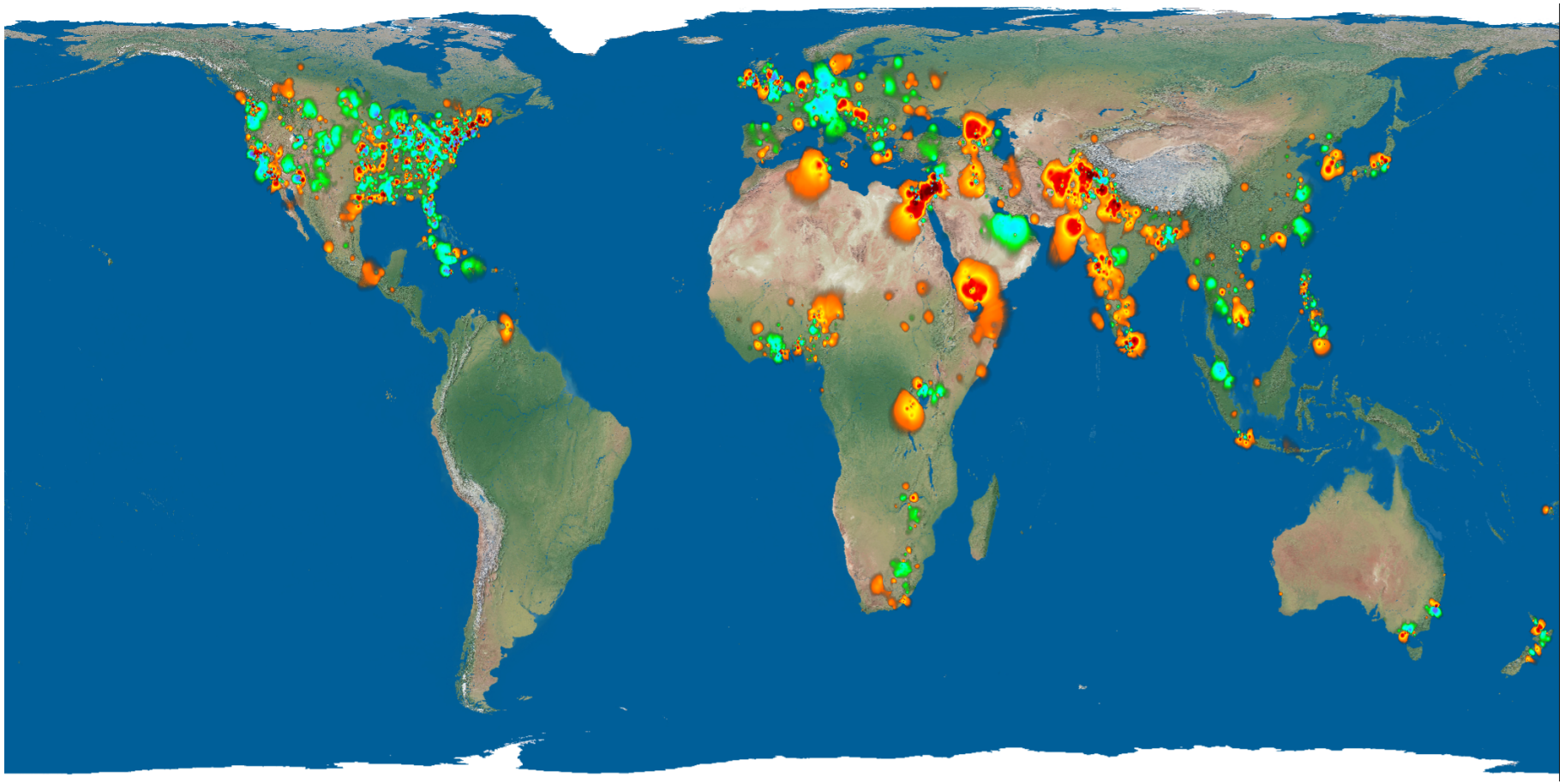


Figure 13: GDELT for 29 January 2011 based on number of events and Goldstein scores

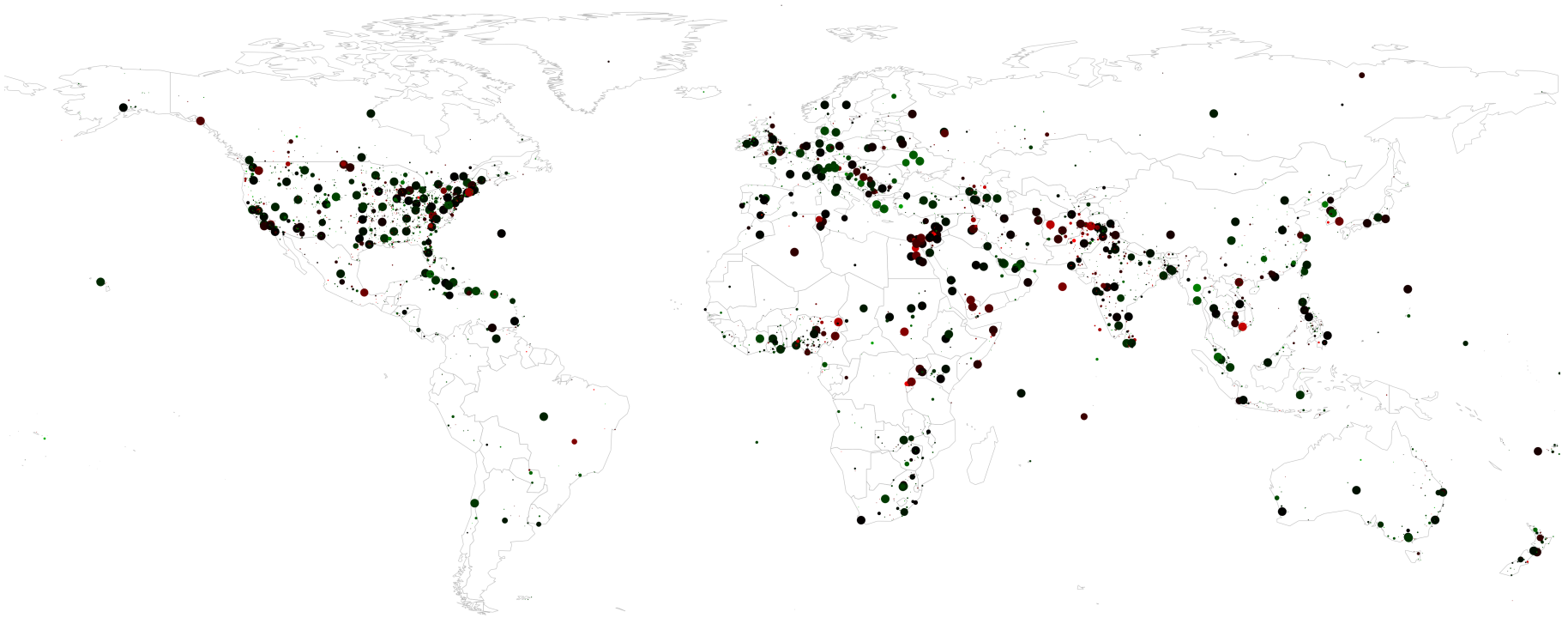


Figure 14: GDELT map for 29 January 2011 based on number of mentions of event and Goldstein scores

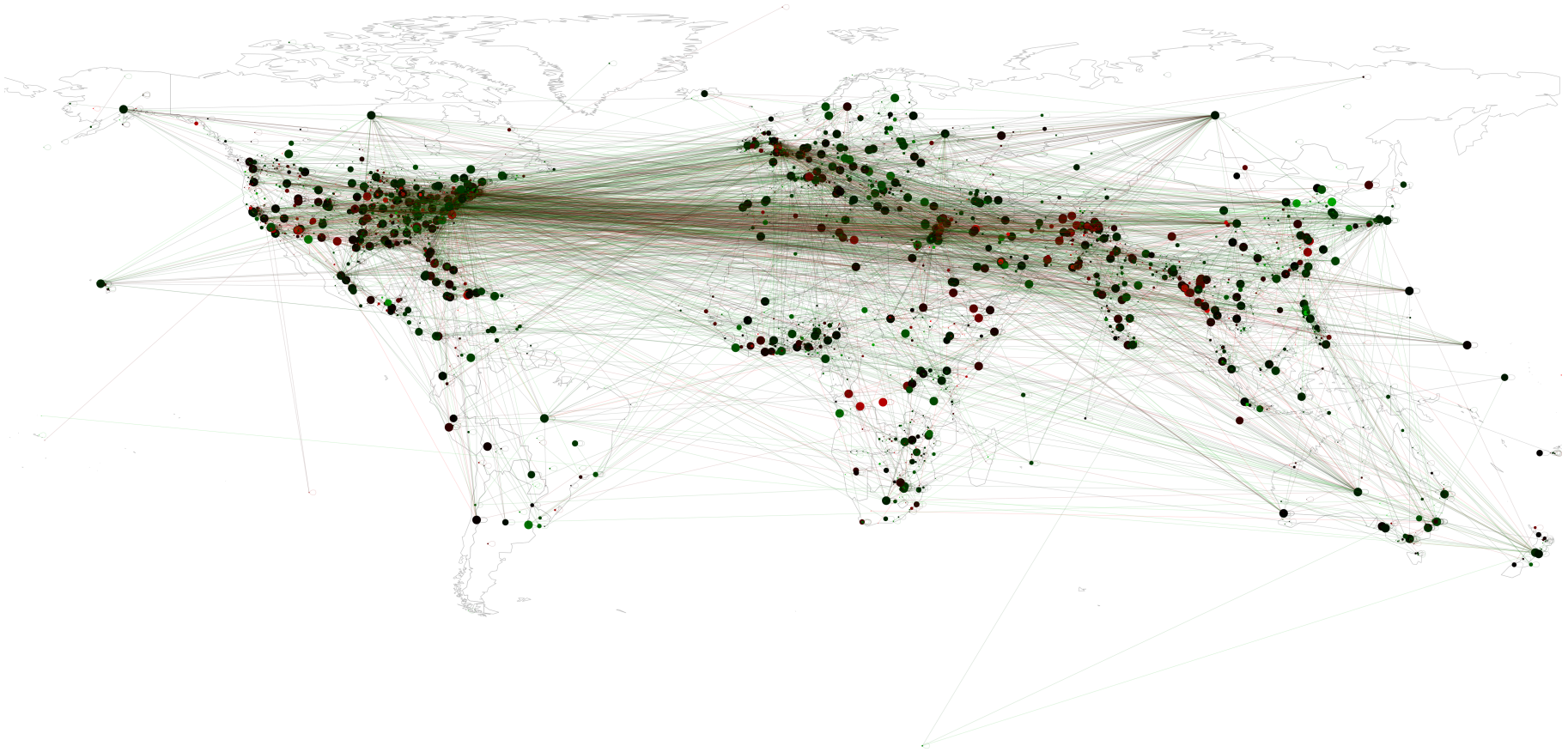


Figure 15: GDELT network map for 29 January 2011 based on dyadic links and average Goldstein scores

5 Comparison to Other Data Sets

5.1 KEDS/Reuters Levant

The KEDS Levant data set (`http://eventdata.psu.edu/data.dir/levant.html`), which covers April-1979 to December-2011, is the only available event data set that is comparable to GDELT in terms of temporal coverage. While it is also coded with TABARI, the dictionaries for this set were extensively configured over a period of years under a number of NSF-funded projects from about 1990 to 2005—the TABARI *.verbs* dictionaries used for the global coding were largely developed on Levant texts—whereas GDELT uses generic global *.actors* and *.agents* dictionaries. In this respect, comparing the two datasets is an assessment of whether the GDELT approach using global dictionaries produces results comparable to the laboriously hand-tuned KEDS data.

The larger difference between the two sets, however, is in the sources. The Levant data consists of two single-source datasets, one based in Reuters, the other—covering only 2000 to the present—on AFP.[6] GDELT, of course, uses a much larger set of sources—though it does not directly incorporate Reuters—so any differences between the two sets will be a function of both the dictionaries and the sources. To maximize coverage, we will look at the Reuters series, and focus on the two dyads in the series with the most activity: Israel→Lebanon (LBN) and Israel→Palestine (PSE).

Figures 16 and 17 show scatterplots for the entire period for two of the quad counts: verbal cooperation and material conflict. Note that the horizontal scale for GDELT is 10-times that of the vertical scale for Reuters: unsurprisingly, the multiple news sources used by GDELT produce, on average, a far higher density of data. The correlations between the two sets are reasonably high, though the scatters are quite wide.

As we saw in Figure 1, the density of data in GDELT varies substantially over time, and shows an exponential increase after 2002 or thereabouts. Furthermore, the activity within these dyads has also changed, particularly with the two *intifadas* and various abortive peace processes in the Israel→Palestine dyad. Consequently, the relationship between the two data set could plausibly vary with time, so we analyzed the three intervals shown in Table 6 separately.

[6]Consult the internal documentation of these datasets for further information on search terms, filtering and so forth. As with GDELT, the Levant data use a "One-A-Day" filter, though unlike GDELT, this is based only on the dyad, not the event location.

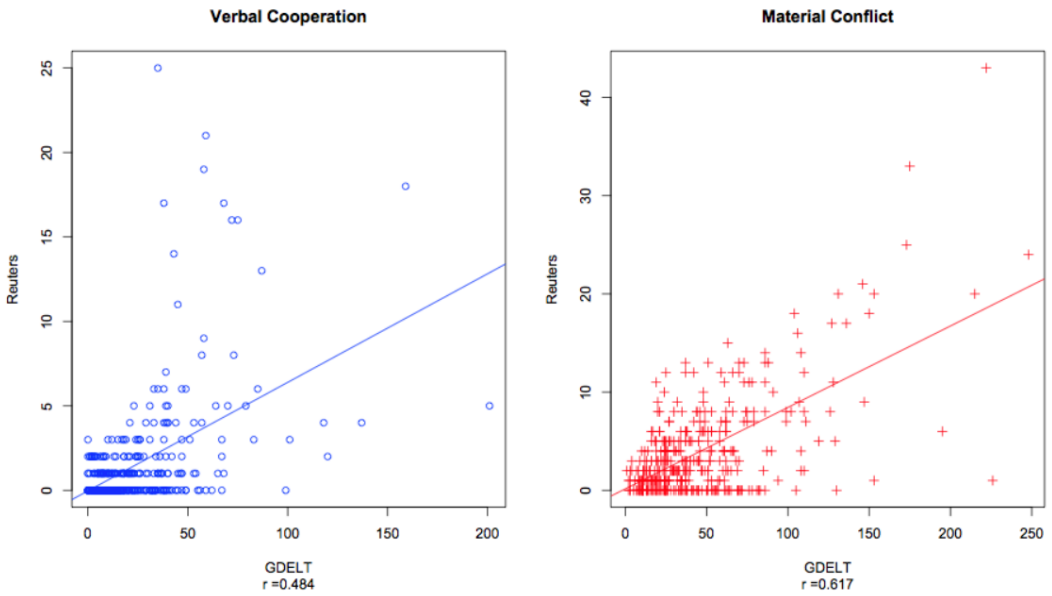


Figure 16: Comparison of KEDS/Reuters and GDELT Quadcounts for Israel → Lebanon

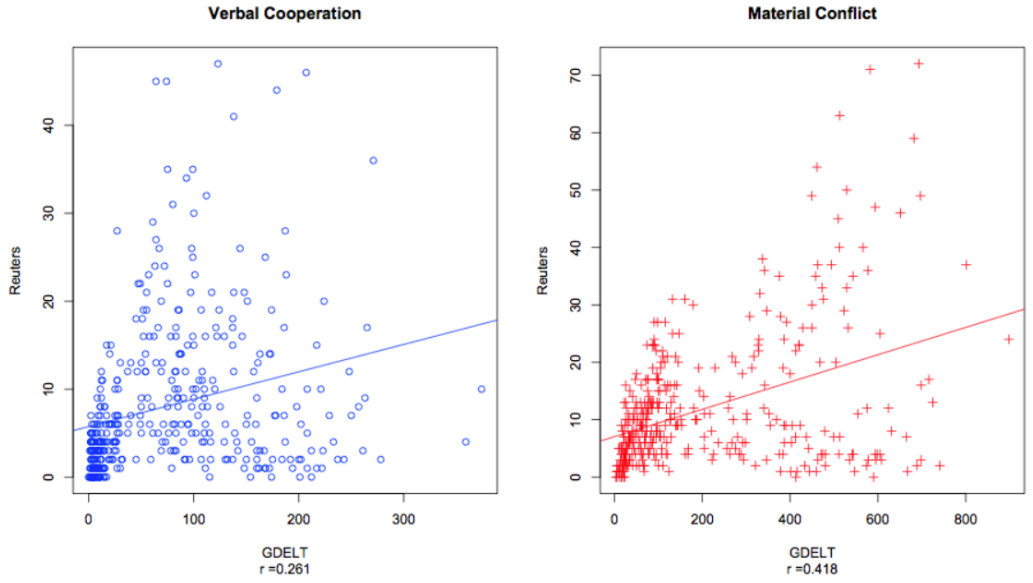
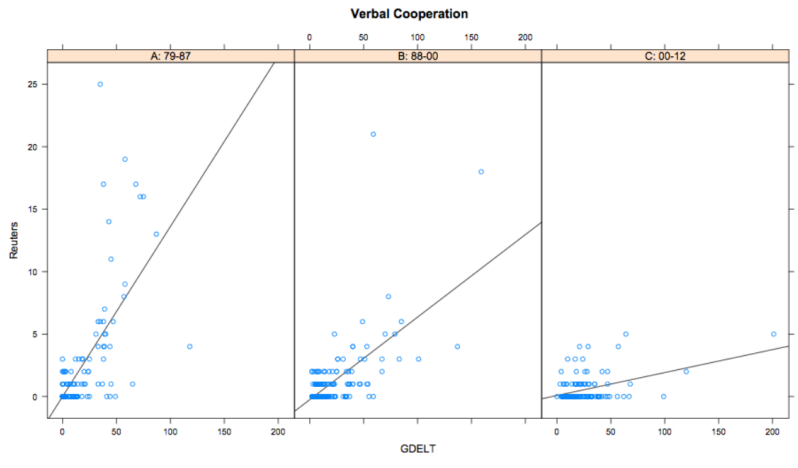
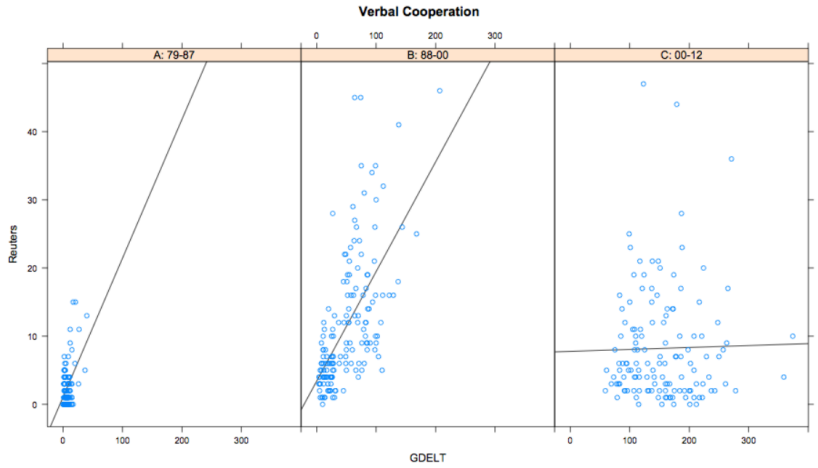
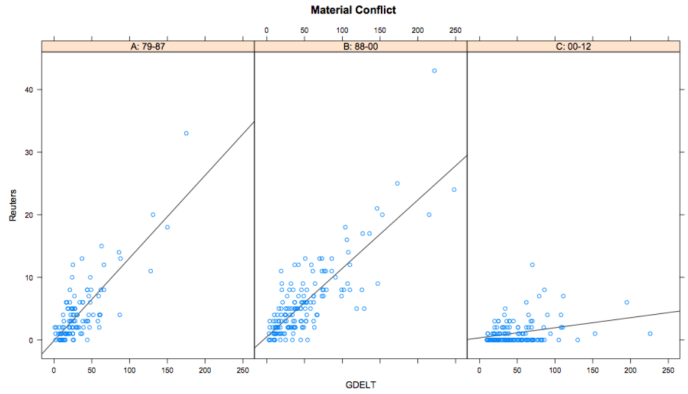
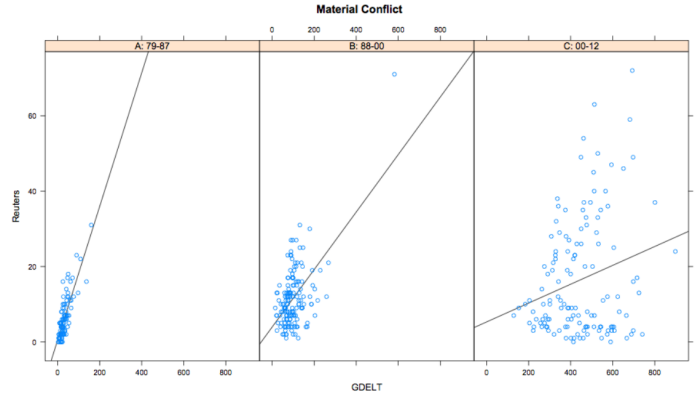


Figure 17: Comparison of KEDS/Reuters and GDELT Quadcounts for Israel → Palestine

Table 6: Hypothesized intervals in Levant data

Interval	Years	ISR → LBN	ISR → PSE
A	1979-1987	Israeli invasion	Camp David
B	1988-1999	Israel-Hizbollah conflict	First intifada, Oslo
C	2000-2011	Cold peace	Second intifada, PA/Hamas split

Table 7: Comparison of KEDS/Reuters and GDELT Quadcounts by Time Period

Israel → Lebanon	Israel → Palestine
	
r = 0.639 r = 0.631 r = 0.420	r = 0.450 r = 0.628 r = 0.020
	
r = 0.794 r = 0.812 r = 0.289	r = 0.812 r = 0.570 r = 0.234

These results are shown in the figures in Table 7—once again note that the GDELT and Reuters counts are on different scales—which shows a very clear pattern of low correlations—nearly zero in the case of Israel → Palestine verbal cooperation—in the post-2000 period, but relatively high correlations pre-2000. In several cases, in fact, the pre-2000 correlations probably exceed those that would be expected from human inter-coder reliability, but in all cases are reasonably high. As emphasized in Figure 18, the low correlations are strongly influenced by a large number of points where GDELT produces events but the Reuters set does not. As far as we can tell, the differences between the sets are mostly due to the differences in the presence of events, rather than differences in what is being reported, though we have not done sufficient tests to rule this out entirely.

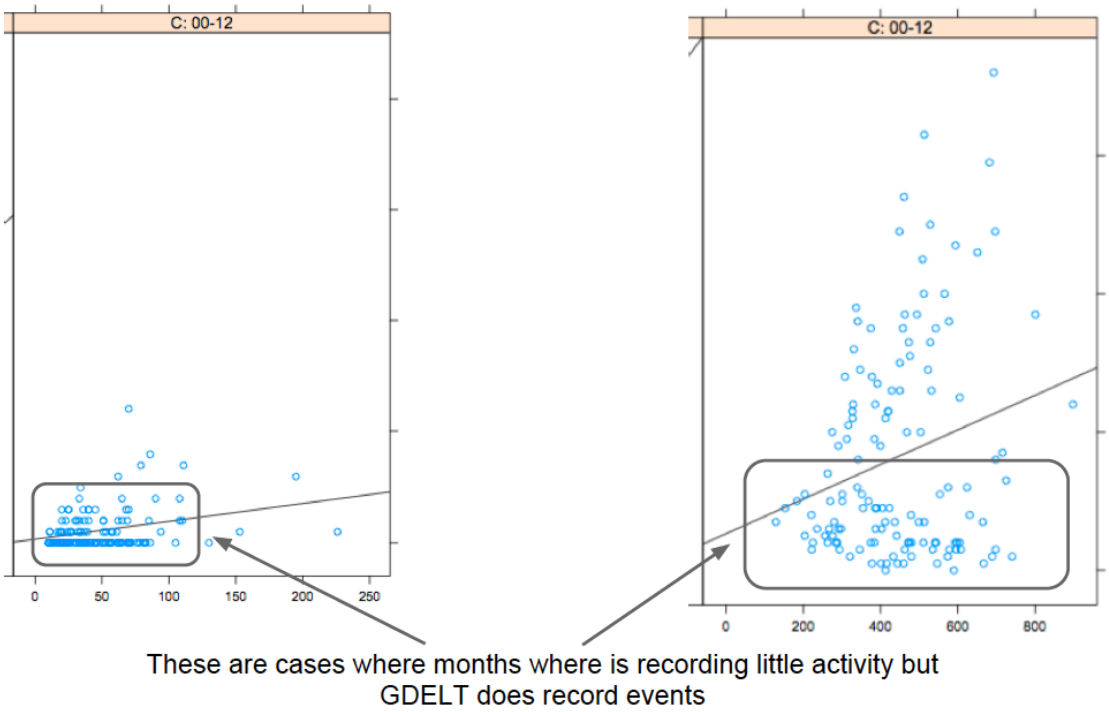


Figure 18: Comparison of KEDS/Reuters and GDELT Quadcounts: 2000-2012 is exceptional

This comparison—along with Figure 1—illustrates what is clearly going to be a major challenge in using GDELT for long time-series studies: the post-2000 increase in the frequency of reports. As shown in Table Table 5, this does not affect all dyads equally: except for a spike in 2006 that corresponds to the 34-day war involving Israel, Lebanon, and Hezbollah in the ISR→LBN data (upper right figure in the table), the period after 2000 for this dyad doesn't look all the different than the period prior to 2000, and the drop-off in the correlations in Table 7 are not nearly as high for Lebanon as for Palestine. Nonetheless, caution is very

much advised on this issue, and given that the international news reporting environment is still changing, it is likely to be some time before this situation stabilizes.

5.2 ICEWS

A second comparison will be with one of the later ICEWS data sets. This was labeled "Release 28" and was coded either with JABARI or possibly some version of JABARI-NLP ; this would have been one of the last versions of the Asian ICEWS data before the project switched to development of the global W-ICEWS set. The data nominally go to Mar-2011 but the last couple of months have very low counts and may have been incomplete, so the series, which begins in Jan-1998, was truncated at Dec-2010. ICEWS is based on a large number of sources from Factiva.

While the ICEWS forecasts primarily focus on internal conflict, the coding dictionaries work for international interactions as well—in fact "international crisis" is one of the ICEWS "events of interest" indicators—and so we will compare the two data sets on four international dyads: China→Taiwan, India→Pakistan, South Korea→North Korea and USA→Japan.

Figure 19 shows the China→Taiwan comparison. The correlations—[0.024 , -0.16, 0.620, 0.275] in the order [VERCP, MATCP, VERCF, MATCF]—are relatively low compared to the KEDS/GDELT comparison, except for the VERCF counts. As with the KEDS/GDELT comparison, GDELT has a higher density of events, though it tends to be only two to three times higher. VERCF is again the exception to this, and the two datasets have roughly equal densities here.

These are not particularly good correlations. Three factors may be contributing to the divergence. First, most of the GDELT sequence being compared here is in the post-2000 period when GDELT is experiencing an exponential increase in density; as we saw in Section 5.1, this period had *much* lower correlations in the KEDS/GDELT comparison. Second, GDELT includes Xinhua, which ICEWS does not include, and Xinhua may be throwing off the totals when compared to the international sources. This in particular might explain why the VERCF (verbal conflict) indicator has the highest correlation: that would be consistent with Xinhua being used as a tool of the Chinese government's generally belligerent foreign policy towards Taiwan, and those policy pronouncements, in turn, would be monitored by the international media.

Finally, there appears to be serious discontinuity in the latter part of the ICEWS sequence, which drops from reporting tens of events per month to ones of events. Using $VERCP >=$

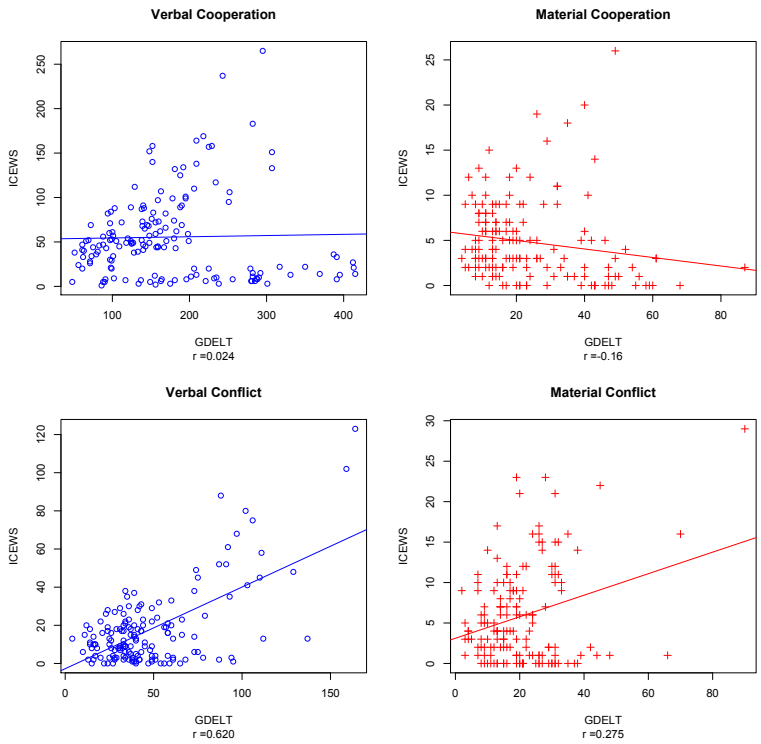


Figure 19: Comparison of ICEWS and GDELT Quadcounts: China→Taiwan 1998-2010

30—that is, an average of at least one event per day—as a threshold, the correlations improve considerably, as shown in Figure 20. The vector of correlations here is [0.525, 0.350., 0.804, 0.583], which is more or less in line with the KEDS/GDELT comparison for the pre-2000 period.

Figures 21, 22 and 23 show scattergrams for the remaining dyads: these generally show patterns similar to those seen in the China→Taiwan case. Figure 21 for India→Pakistan shows the same pattern of a high correlation [0.623] for VERCP—though this is clearly inflated by an outlying point which is similar in both data sets—and relatively low correlations [0.16 to 0.25] for the remaining counts; the ratio of the GDELT to ICEWS counts is again in the range of two to three.[7] South Korea→North Korea, Figure 22, has higher correlations, in the range [0.55-0.75] except for MATCP, though again these are inflated by outliers. The ratio of GDELT to ICEWS counts is substantially higher here, in the range of five to ten; again it is possible that Xinhua accounts for the difference. Finally, the USA→Japan dyad, Figure 23, has very low correlations, once again strongly influenced by the very low counts on all of the ICEWS indicators except VERCP.

These comparisons clearly need to be explored in further detail, the most critical issue being

[7]We experimented with eliminating low frequency VERCP cases here and it did not make much difference.

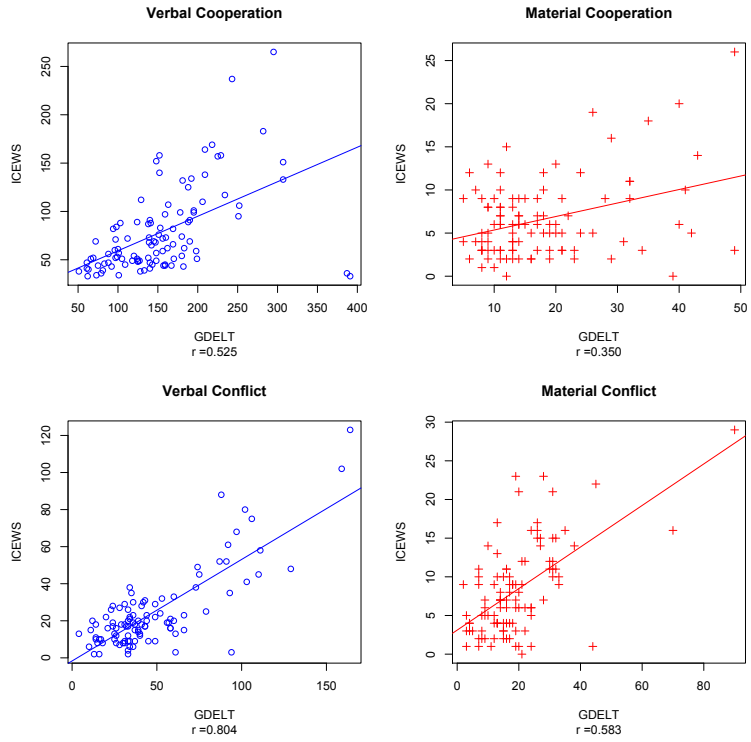


Figure 20: Comparison of ICEWS and GDELT Quadcounts: China→Taiwan 1998-2010 with $VERCP < 30$ cases removed

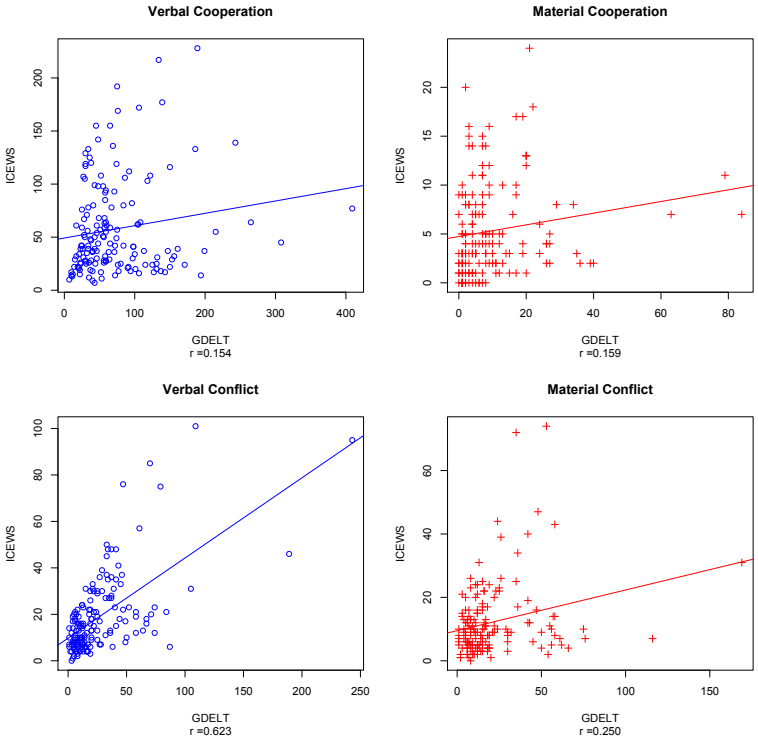


Figure 21: Comparison of ICEWS and GDELT Quadcounts: India→Pakistan 1998-2010

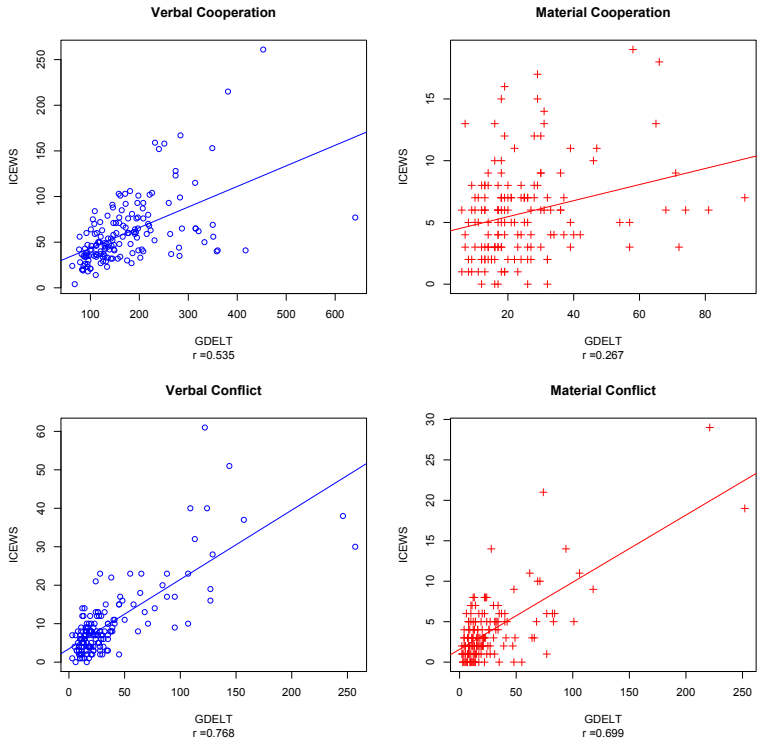


Figure 22: Comparison of ICEWS and GDELT Quadcounts: South Korea→North Korea 1998-2010

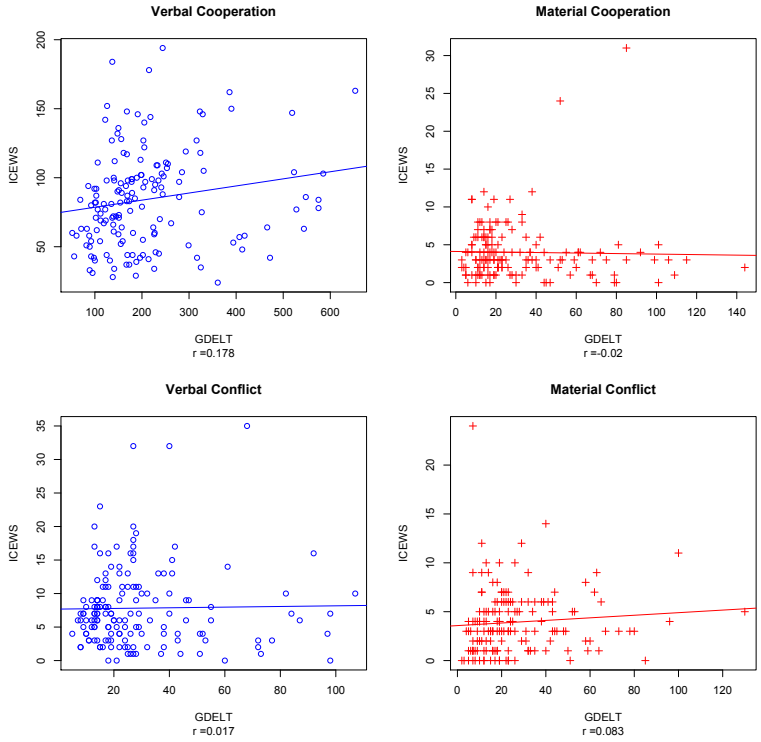


Figure 23: Comparison of ICEWS and GDELT Quadcounts: USA→Japan 1998-2010

further analysis to ascertain whether the difference in the counts is due to ICEWS being more selective—at various points, the project was working on making the coding very sensitive to false positives, particularly on conflict events—or whether GDELT is capturing more detail, particularly in the post-2000 period, because it is using a wider variety of web-based sources.[8] A systemic drop in reports in ICEWS at the same time GDELT is experiencing an exponential increase in reports would, of course, be a perfect storm for the sequences not correlating.

To further explore this possibility, we ran two-group t-tests on the 32 sequences (quad categories x dyad x GDELT/ICEWS), splitting the series at Jan-2005. Twelve of the sixteen tests showed significant differences ($p < 0.05$) for ICEWS, and the same ratio occurred for GDELT. However, the *direction* of these were quite different between the two data sets: of the twelve significant differences in ICEWS, ten were positive (the counts in the pre-2005 were higher), whereas for GDELT only four were positive. Finally, except in the South Korea→North Korea (where all directions were the same) in all of the cases where both t-tests were significant, the changes were in opposite directions. All of this would suggest that much of the difference between the sets is due to changes in the baseline frequencies rather than the coding of specific events.

All factors being equal (though in this case, they clearly aren't equal due to differences in the source texts), the ICEWS data should be more accurate than the GDELT data because Jabari-NLP has some clear advantages over Tabari [Schrodt and Van Brackle, 2013]. Our sense, however, is that there is more going on here than just the difference in coding engines, since those accuracy improvements are unlikely to have exclusively resulted in the elimination of events. At this point, it would be very helpful if we had additional updated and documented sets of ICEWS and W-ICEWS data, but at present these do not seem to be available. Until we further explore the characteristics of the ICEWS data, this comparison should not be considered definitive.[9]

6 Future Plans and Collaboration Opportunities

The authors are already well underway on work for the next release of this event dataset, which include an expanded event taxonomy, an array of new event attributes (including

[8] An additional factor that might accounts for some of the differences is that fact that GDELT is using *location-based* duplicate filtering—events are considered duplicates only if they are the same event and occur in the same city—whereas as far as we know, ICEWS was using only *dyad-based* duplicate filtering.

[9] "Not definitive": yes, that means you, BBN and MITRE...

estimates of the number killed and injured), and a completely new event processing system based a flexible new natural language processing system. In addition, the next release will extend coverage back to 1800 for all countries.

As noted in Section 3.1.2, we have new versions of the verb dictionaries that are organized around WordNet synonym sets for both the core verbs and common noun sets such as those using units of currency. These should make the dictionaries more robust, since a general phrase such as "Nation X has agreed to provide Nation Y *currency*-million in development aid" will need to be entered only once, rather than separately for each currency. All the work has been done on these; we just need to do some additional formatting.

While TABARI now has the ability to handle such "synsets," Schrodt along with Michael Ward and Jay Ulfelder recently received NSF-funding which will provide for the development of a completely new coder written in Python, at which point we will retire TABARI . This programming, which will be developed on the open-collaboration platform GitHub, was motivated by several issues

- While thoroughly debugged—TABARI coded the 200-million events of GDELT without crashing—the code base is about twelve years old and written in the computer language C++, which no longer has a large (or young) programming community;[10]
- Python is *much* more suited than C++ for text processing—we are hoping the choice of Python will reduce the remaining required programming to a point where this will be a summer-length project—and in a number of experiments we have done with very large scale textual databases, seems sufficiently fast (and in any case, we can always run large jobs on a cluster computer)
- Based on the experience of the Lockheed JABARI-NLP system developed for the latter phases of ICEWS [Schrodt and Van Brackle, 2013], we will be incorporating a number of open-source natural language preprocessors so that the system can work on a parsed representation of the text, rather than continuing the "shallow parsing" approach of TABARI and KEDS. This should be particularly helpful in improving the accuracy of target actor identification.
- The various GDELT pre-processing steps described in Section 3.3.1 will be also incorporated into a Python-based pre-processing suite which can be used for standard processing.

[10]There also appear to be some problems getting the C++ "ncurses" library, which TABARI uses for its interface, to run on some versions of Ubuntu Linux.

We are also hoping to leverage efforts elsewhere in the social science data community for the further standardization of non-state actor codes: while we have developed *a* set of religious and ethnic codes in CAMEO, one of the primary advantages of automated coding is that recoding using a new (or multiple) systems is quite easy once dictionaries have been developed, whereas this was nearly impossible, and we are certainly open to other suggestions. In addition, we will be extending the CAMEO coding system—and the dictionaries—to incorporate new categories of events relating to finance, criminal activity, disease, and natural disasters. Finally, under the new NSF funding we are expecting to set up one or more "crowd-sourcing" platforms—most likely within expert communities (and their conscripted students) rather than for web denizens in general—that will allow collaborative development of dictionaries, particularly in maintaining up-to-date lists of political actors, and spot-checking of the accuracy of the coding.

If these changes proceed on schedule, we are hoping that around September 2013 or thereabouts we should be in a position to do a complete GDELT 2.0 recode. In the meantime, however, we will first deploy the existing data in a mySQL database on a high-bandwidth server at the University of Texas at Dallas (with the assistance of Patrick Brandt) and implement the system for near-real-time coding with daily updates, which has been running in an experimental mode since October 2012. Updates on the status of these systems can be found at `http://eventdata.psu.edu/data.dir/GDELT.html`, which also contains a "reduced" version of the data set that has only the core event, actor and location variables (but for all countries and years). When the system is fully operational, we will broadcast announcements on various international relations and methodology listservs.

7 Appendix: GDELT CODEBOOK Version 1.0

Documentation Version Date: 25-July-2012

7.1 INTRODUCTION

This codebook provides a quick overview of the fields in the GDELT data file format and their descriptions. GDELT event records are in the dyadic CAMEO format, capturing two actors and the action performed by Actor1 upon Actor2. A wide array of variables break out the raw CAMEO actor codes into their respective fields to make it easier to interact with the data, the Action codes are broken out into their hierarchy, the Goldstein ranking score is provided, an average tone score is provided for all coverage of the event, several indicators of importance are provided, and a special array of georeferencing fields offer estimated landmark-centroid-level geographic positioning of both actors and the location of the action.

A reduced version of the data set which contains the basic event, actor and geolocation variables for all countries for Jan-1979 to June-2012 can be downloaded from a link at `http://eventdata.psu.edu/data.dir/GDELT.html`. The file is about 650Mb compressed and includes Python programs for doing basic subsetting and generating counts and R graphics: the file "GDELT.reduced.documentation.txt" describes the file format and utility programs; files are in Unix format.

We are in the process of installing GDELT on a server at UT/Dallas with appropriate bandwidth; this installation will provide a mySQL facility for subsetting and automatically update the dataset on a daily basis.

7.2 BASE ATTRIBUTES

These attributes capture the date and raw actor codes for Actor1 and Actor2.

- **Day**. Date the event took place in YYYYMMDD format.
- **GlobalEventID**. Globally unique identifier assigned to each event record that uniquely identifies it in the master dataset. NOTE: While these will often be sequential with date, this is NOT always the case and this field should NOT be used to sort events by date: the date fields should be used for this.

- **MonthYear**. Alternative formatting of the event date, in YYYYMM format.
- **Year**. Alternative formatting of the event date, in YYYY format.
- **FractionDate**. Alternative formatting of the event date, computed as YYYY.FFFF, where FFFF is the percentage of the year completed by that day. This collapses the month and day into a fractional range from 0 to 0.9999, capturing the 365 days of the year. The fractional component (FFFF) is computed as (MONTH * 30 + DAY) / 365. This is an approximation and does not correctly take into account the differing numbers of days in each month or leap years, but offers a simple single-number sorting mechanism for applications that wish to estimate the rough temporal distance between dates.
- **Actor1Code**. The complete raw CAMEO code for Actor1 (includes geographic, class, ethnic, religious, and type classes). May be blank if the system was unable to identify an Actor1.
- **Actor1Name**. The actual name of the Actor 1. In the case of a political leader or organization , this will be the leaders formal name (GEORGE W BUSH, UNITED NATIONS), for a geographic match it will be either the country or capital/major city name (UNITED STATES / PARIS), and for ethnic, religious, and type matches it will reflect the root match class (KURD, CATHOLIC, POLICE OFFICER, etc). May be blank if the system was unable to identify an Actor1.
- **Actor2Code**. The complete raw CAMEO code for Actor2 (includes geographic, class, ethnic, religious, and type classes). May be blank if the system was unable to identify an Actor2.
- **Actor2Name**. The actual name of the Actor 2. In the case of a political leader or organization , this will be the leaders formal name (GEORGE W BUSH, UNITED NATIONS), for a geographic match it will be either the country or capital/major city name (UNITED STATES / PARIS), and for ethnic, religious, and type matches it will reflect the root match class (KURD, CATHOLIC, POLICE OFFICER, etc). May be blank if the system was unable to identify an Actor2.

7.3 ACTOR CODE BREAKOUT

The Actor1 and Actor2 fields may contain multiple codes indicating geographic, ethnic, and religious affiliation and the actors role in the environment (political elite, military officer,

rebel, etc). These codes may be combined in any order, and are encoded in a single character field consisting of a string of concatenated 3-digit codes. To make it easier to utilize this information in analysis, this section breaks these codes out into a set of individual columns.

- **Actor1CountryCode**. The 3-digit CAMEO code for the country affiliation of Actor1.
- **Actor1CountryLabel**. The human-readable name of the country affiliation of Actor1.
- **Actor1KnownGroupCode**. If Actor1 is a known IGO/NGO/rebel organization (al-Qaeda, United Nations, World Bank, etc) with its own CAMEO code, this field will contain that code.
- **Actor1KnownGroupLabel**. The human-readable formal name for Actor1KnownGroupCode.
- **Actor1EthnicCode**. If the source document specifies the ethnic affiliation of Actor1 and that ethnic group has a CAMEO entry, the CAMEO code is entered here. NOTE: a few special groups like ARAB may also have entries in the type column due to legacy CAMEO behavior.
- **Actor1EthnicLabel**. The human-readable formal name for Actor1EthnicCode.
- **Actor1Religion1Code**. If the source document specifies the religious affiliation of Actor1 and that religious group has a CAMEO entry, the CAMEO code is entered here. NOTE: a few special groups like JEW may also have entries in the geographic or type columns due to legacy CAMEO behavior.
- **Actor1Religion1Label**. The human-readable formal name for Actor1Religion1Code.
- **Actor1Religion2Code**. If multiple religious codes are specified for Actor1, this contains the secondary code. Some religion entries automatically use two codes, such as Catholic, which invokes Christianity as Code1 and Catholicism as Code2.
- **Actor1Religion2Label**. The human-readable formal name for Actor1Religion2Code.
- **Actor1Type1Code**. The 3-digit CAMEO code of the CAMEO type or role of Actor1, if specified. This can be a specific role such as Police Forces, Government, Military, Political Opposition, Rebels, etc, a broad role class such as Education, Elites, Media, Refugees, or organizational classes like Non-Governmental Movement. Special codes such as Moderate and Radical may refer to the operational strategy of a group.
- **Actor1Type1Label**. The human-readable formal name for Actor1Type1Code.

- **Actor1Type2Code**. If multiple type/role codes are specified for Actor1, this returns the second code.
- **Actor1Type2Label**. The human-readable formal name for Actor1Type2Code.
- **Actor1Type3Code**. If multiple type/role codes are specified for Actor1, this returns the third code.
- **Actor1Type3Label**. The human-readable formal name for Actor1Type3Code.

These codes are repeated for Actor2, using the prefix Actor2 instead of Actor1. As with Actor1, if no Actor2 could be extracted, these fields will be blank. Only in extremely rare circumstances will both Actor1 and Actor2 be blank.

7.4 EVENT ACTION ATTRIBUTES

These fields break out various attributes of the event action and offer several mechanisms for assessing the importance or immediate-term impact of an event.

- **IsRootEvent**. The system codes every event across an entire document, using an array of techniques to deference and link information together. A number of previous projects such as the ICEWS initiative have found that events occurring in the lead paragraph of a document tend to be the most important and are the least likely to have any errors. Thus, this flag can be used as a proxy for the rough importance of an event to create subsets of the event stream.
- **EventCode**. This is the raw CAMEO action code describing the action that Actor1 performed upon Actor2.
- **EventDesc**. This is the human-readable formal label for the given CAMEO action code.
- **EventBaseCode**. CAMEO event codes are defined in a three-level taxonomy. For events at level three in the taxonomy, this yields its level two leaf root node. For example, code 0251 (Appeal for easing of administrative sanctions) would yield an EventBaseCode of 025 (Appeal to yield). This makes it possible to aggregate events at various resolutions of specificity. For events at levels two or one, this field will be set to EventCode.
- **EventBaseDesc**. This is the human-readable formal label for EventBaseCode.

- **EventRootCode**. Similar to EventBaseCode, this defines the root-level category the event code falls under. For example, code 0251 (Appeal for easing of administrative sanctions) has a root code of 02 (Appeal). This makes it possible to aggregate events at various resolutions of specificity. For events at levels two or one, this field will be set to EventCode.

- **EventRootDesc**. This is the human-readable formal label for EventBaseCode.

- **QuadClass**. The entire CAMEO event taxonomy is ultimately broken into four primary classifications: Verbal Cooperation, Material Cooperation, Verbal Conflict, and Material Conflict. This field specifies this primary classification for the event type, allowing analysis at the highest level of aggregation.

- **GoldsteinScale**. Each CAMEO event code is assigned a numeric score from -10 to +10, capturing the likely impact that type of event will have on the stability of a country. This is known as the Goldstein Scale. This field specifies the Goldstein score for each event type. NOTE that this score is based on the type of event, not the specifics of the actual event record being recorded thus two riots, one with 10 people and one with 10,000, will both receive the same Goldstein score. This can be aggregated to various levels of time resolution to yield an approximation of the stability of a geography over time.

- **NumMentions**. This is the total number of mentions of this event across all source documents. Multiple references to an event within a single document also contribute to this count. This can be used as a method of assessing the importance of an event: the more discussion of that event, the more likely it is to be significant. The total universe of source documents and the density of events within them vary over time, so it is recommended that this field be normalized by the average or other measure of the universe of events during the time period of interest.

- **NumSources**. This is the total number of information sources containing one or more mentions of this event. This can be used as a method of assessing the importance of an event: the more discussion of that event, the more likely it is to be significant. The total universe of sources varies over time, so it is recommended that this field be normalized by the average or other measure of the universe of events during the time period of interest.

- **NumArticles**. This is the total number of source documents containing one or more mentions of this event. This can be used as a method of assessing the importance of an event: the more discussion of that event, the more likely it is to be significant. The

total universe of source documents varies over time, so it is recommended that this field be normalized by the average or other measure of the universe of events during the time period of interest.

- **AvgTone**. This is the average tone of all documents containing one or more mentions of this event. The score ranges from -100 (extremely negative) to +100 (extremely positive). Common values range between -10 and +10, with 0 indicating neutral. This can be used as a method of filtering the context of events as a subtle measure of the importance of an event and as a proxy for the impact of that event. For example, a riot event with a slightly negative average tone is likely to have been a minor occurrence, whereas if it had an extremely negative average tone, it suggests a far more serious occurrence. A riot with a positive score likely suggests a very minor occurrence described in the context of a more positive narrative (such as a report of an attack occurring in a discussion of improving conditions on the ground in a country and how the number of attacks per day has been greatly reduced).

7.5 EVENT GEOGRAPHY

The final set of fields add a novel enhancement to the CAMEO taxonomy, georeferencing each event along three primary dimensions to the landmark-centroid level. To do this, the fulltext of the source document is processed using fulltext geocoding and automatic disambiguation to identify every geographic reference. The closest reference to each of the two actors and to the action reference are then encoded in these fields. The georeferenced location for an actor may not always match the Actor1CountryCode field, such as in a case where the President of Russia is visiting Washington, DC in the United States, in which case the Actor1CountryCode would contain the code for Russia, while the georeferencing fields below would contain a match for Washington, DC. It may not always be possible for the system to locate a match for each actor or location, in which case one or more of the fields may be blank. Finally, the Action fields capture the location information closest to the point in the event description that contains the actual statement of action.

- **Actor1Geo_Type**. This field specifies the geographic resolution of the match type and holds one of the following values: COUNTRY (the match was at the country level), USSTATE (the match was to a US state), USLOC (the match was to a US city or landmark), WORLDLOC (the match was to a city or landmark outside the US). This can be used to filter events by geographic specificity, for example, extracting only those events with a landmark-level geographic resolution for mapping. Note that both

COUNTRY and USSTATE matches will still provide a latitude/longitude pair, which will be the centroid of that country or state. Matches to foreign Administrative Division 1s (ADM1s) (the rough equivalent of a US state) will be coded as a WORLDLOC location.

- **Actor1_Geo_Fullname**. This is the full human-readable name of the matched location. In the case of a country it is simply the country name. For US states it is in the format of State, United States, while for all other matches it is in the format of Landmark, State/ADM1, Country. This can be used to label locations when placing events on a map.
- **Actor1Geo_CountryCode**. This is the 2-character FIPS10-4 country code for the location.
- **Actor1Geo_ADM1Code**. This is the 2-character FIPS10-4 administrative division 1 (ADM1) code for the administrative division housing the landmark. In the case of the United States, this is the 2-character shortform of the states name (such as TX for Texas).
- **Actor1Geo_Lat**. This is the centroid latitude of the landmark for mapping.
- **Actor1Geo_Long**. This is the centroid longitude of the landmark for mapping.

These codes are repeated for Actor2 and Action, using those prefixes.

References

Edward E. Azar. The conflict and peace data bank (COPDAB) project. *Journal of Conflict Resolution*, 24:143–152, 1980.

Roy F. Baumeister and John Tierney. *Willpower: Rediscovering the Greatest Human Strength*. Penguin, New York, 2011.

Doug Bond, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles L. Taylor. Integrated data for events analysis (IDEA): An event typology for automated events data development. *Journal of Peace Research*, 40(6):733–745, 2003.

Deborah J. Gerner, Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. The machine coding of events from regional and international sources. *International Studies Quarterly*, 38:91–119, 1994.

Deborah J. Gerner, Philip A. Schrodt, and Ömür Yilmaz. *Conflict and Mediation Event Observations (CAMEO) Codebook*. `http://eventdata.psu.edu/data.dir/cameo.html`, 2009.

Joshua S. Goldstein. A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution*, 36:369–385, 1992.

Gary King and Will Lowe. 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, 2004.

Kalev Leetaru. Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched wikipedia. *D-Lib*, 18(9/10), September/Octoer 2012.

Kalev Leetaru. *Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space.* PhD thesis, University of Illinois, 2013.

Russell J Leng. *Behavioral Correlates of War, 1816-1975. (ICPSR 8606)*. Inter-University Consortium for Political and Social Research, Ann Arbor, 1987.

Charles A. McClelland. *World Event/Interaction Survey Codebook (ICPSR 5211)*. Inter-University Consortium for Political and Social Research, Ann Arbor, 1976.

Patrick McGowan, Harvey Starr, Gretchen Hower, Richard L. Merrit, and Dina A. Zinnes. International data as a national resource. *International Interactions*, 14:101–113, 1988.

Richard L. Merritt, Robert G. Muncaster, and Dina A. Zinnes, editors. *International Event Data Developments: DDIR Phase II*. University of Michigan Press, Ann Arbor, 1993.

Slava Mikhaylov, Michael Laver, and Kenneth Benoit. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, 20(1):78–91, 2012.

Andrea Ruggeri, Theodora-Ismene Gizelis, and Han Dorussen. Events data as Bismarck's sausages? intercoder reliability, coders' selection, and data quality. *International Interactions*, 37(1):340–361, 2011.

Philip A. Schrodt and Deborah J. Gerner. Validity assessment of a machine-coded event data set for the Middle East, 1982-1992. *American Journal of Political Science*, 38:825–854, 1994.

Philip A. Schrodt and David Van Brackle. Automated coding of political event data. In V.S. Subrahmanian, editor, *Handbook of Computational Approaches to Counterterrorism.*, pages 23–50. Springer, 2013.

E. Shook, K. Leetaru, G. Cao, A. Padmanabhan, and S Wang. Happy or not: Generating topic-based emotional heatmaps for Culturomics using CyberGIS. In *IEEE 8th International Conference on EScience*, pages 1–6. IEEE, 2012.

Jay Yonamine. *A Nuanced Study of Political Conflict Using the Global Dataset of Events Location And Tone (GDELT) Dataset*. PhD thesis, Pennsylvania State University, 2013.